NEWS Analytics

Alexander Corpstein
James Nieberding

FINAL REPORT

NEWS Analytics

Alexander Corpstein
James Nieberding

CONCEPT OF OPERATIONS

REVISION – 3
03 December 2023

CONCEPT OF OPERATIONS
FOR
News Analytics

TEAM <12>

APPROVED BY:

_____
Project Leader Date

_____
Prof. Kalafatis Date

_____
T/A Date

Change Record

Rev.	Date	Originator	Approvals	Description
1	9/14/2023	James Nieberding		First Draft
2	9/28/2023	Alexander Corpstein		Revised For Midterm Report
3	12/03/2023	Alexander Corpstein		Revised For Final Report
4	4/29/2024	James Nieberding		Revised For End of Project Report

Table of Contents

List of Tables

List of Figures

1. Executive Summary

This project will create a tool that can assess the sentiment and subjectivity of Texas A&M University within the news. The importance of understanding how public perception of a specific institution is currently and how it is changing is vital for making informed decisions. News articles past and present will be accounted for to see trends of the general perception through time. The goals of this tool are to grade an article on its sentiment, subjectivity while also accounting for its time, publisher, topic, and other metadata. Using a trained gradient boosting model, the system will be able to assess the content and trends of the news. Based on the training dataset, the model will create a narrative curve over a predefined narrative with regards to Texas A&M. Peaks in the curve would be normalized, and events combined into clusters which can be presented to show the causes and effects of the highest peak in the narrative curve.

2. Introduction

This project is at the request of the Los Alamos National Laboratory to assess public perception of an institution by analyzing the sentiment and subjectivity of news articles and creating a model based on the data obtained. Public perception of an institution is important for an organization to be cognizant of. Based on the data collected, better decisions can be made for the future of the institution.

2.1. Background

There are existing systems that aggregate new articles into forms that can be searched and read by a user, and there also exists software that provides data analysis on news articles. Examples of these systems include News Crawl by Common Crawl, NewsFetch, and MonkeyLearn. Instead of focusing on general news, this project will provide a specific response and report directly to an institution to aid in understanding the public perception of its organization via a specific model. This will allow an institution to receive a specific technical report based on a predefined model to aid them in decision-making.

2.2. Overview

This project is broken into 4 distinct parts: The data scraper, database storage, the NLP model, and the final narrative curve model. The finalized report is separate from the system itself and will be done by all team members. The data scraper will take a given article from the web and extract the raw text from it. The data will then go into the database for use by the NLP (natural language processor). The NLP is responsible for taking the raw data from the database, cleaning it, and applying a sentiment and subjectivity recognition model. The final model is then responsible for using the sentiment and subjectivity scores and article metadata to create a narrative curve best fit for the general trend of a given topic.

2.3. Referenced Documents and Standards

Document Number	Revision/Release Date	Document Title
1	3.11.5	Python Library Reference
2	a320e5f6	Scrapy 2.10 Documentation
3	3.12.0	SQlite3 Documentation
4	v4.0	Selenium Documentation
5	v3.5.20357	SSHFS-Win
6	096047c5	dmlc XGBoost
7	3.8.1	NLTK Documentation
8	0.16.0.	Textblob Documentation
9	2.1.1	Pandas Documentation
10	2023-07-06	Tensorflow Documentation
11	4.3.0	Gensim Documentation
12	0.21.3	ScikitLearn Documentation
13	1.26	Numpy Documentation

3. Operating Concept

3.1. Scope

This project is focused on the sentiment and subjectivity of news articles about Texas A&M from various sources across the web. This project will allow Los Alamos National Laboratory to better understand the public perception of the institution via analyzing new articles. This will give a better understanding of how public perception can be utilized in the decision-making process of an organization. The methods for data extraction and analysis will be presented so they can be applied to other institutions and applications.

3.2. Operational Description and Constraints

To operate this system, the user needs to be one of the creators of the project or needs to have the source code for the project, given that there is no UI or website for remote access. From a single device, the code is run and output data and graphs are given to the user. There will be no search function, and the data will come from predetermined online news outlets that have been scraped and cleaned from the web.

As such the system is constrained to be used by the project team members. This keeps the program and code to be operated constrained to the project members' devices. The Olympus server will host the database, and all access to the database will be confined to the team members, who will have access to the server.

3.3. System Description

This project can be categorized into 4 parts: An article scraper, a database, an article analyzer, and a narrative structure model.

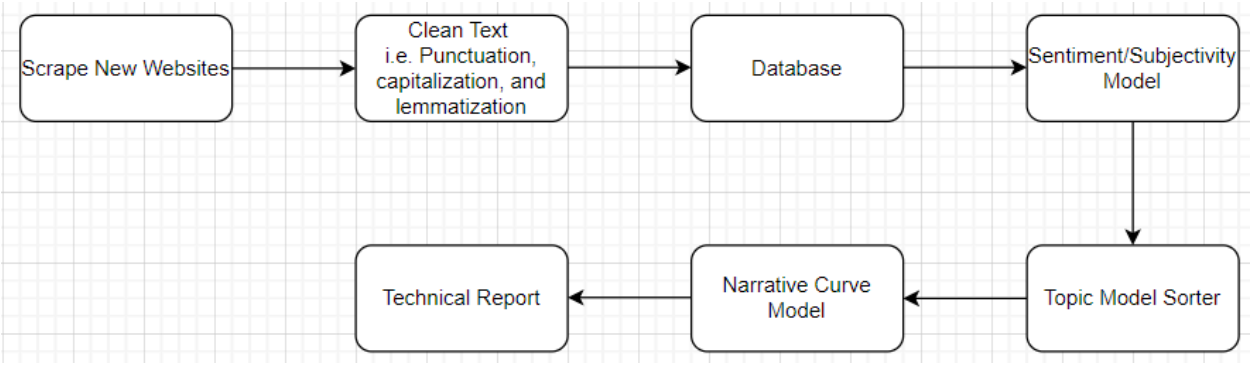


Figure 1: System Block Diagram

Using the Python libraries Scrapy and Selenium, news articles from the internet will be scrapped for content, date, publisher, topic, and other metadata. This data will then be

locally saved for processing and verification before being sent to the SQlite3 database. Utilizing an SQlite3 database for storing the data will allow for ease of access for the sentiment/subjectivity analyzer, emotion recognition model and the narrative structure model.

The sentiment/subjectivity analyzer portion of the project will take the raw article data from the database and clean the text for the NLP analyzer. The NLP libraries that will be used are called Python Natural Language Toolkit (NLTK) and TextBlob. NLTK will be responsible for assigning a score for sentiment from -1 (completely negative) to 1 (completely positive), with 0 as a neutral score. Textblob will be responsible for assigning a subjectivity score for each article from 0 (pure fact) to 1 (pure opinion). Sentiment and subjectivity scores for each article will be passed onto the model portion for further analysis.

The narrative curve model portion of the system is responsible for generating the output and analyzed data which will be used in this final report. Using XGBoost, a gradient-boosting algorithm, the model will learn trends and make predictions based on training datasets that will then be used to assess the article data coming from the analyzer. The model will take in grading data on news articles about a specific topic as well as when it was published to create a narrative curve over time. Using the algorithms ability to create predictive models we can predict the shape of the curve in the future. Gradient boosting has been proven to be effective in many different kinds of applications involving prediction and analytics.

A narrative curve can really be up to interpretation and can be shaped in any way that the narrative is defined. Simply put, a narrative curve is the general trend of a story going beginning with introduction, exposition, climax, and ending with the falling action and resolution. In the real world, most story arcs don't follow curves shown below exactly, however often many peaks have causes and effects which can be grouped together. The goal of this tool is to model the shape of the narrative so we can understand what events led to the highest peak and what are the rippling effects of the peak as time goes on. Once the model is able to develop these curves, we can then work towards training the model to create predictions of a certain curve.

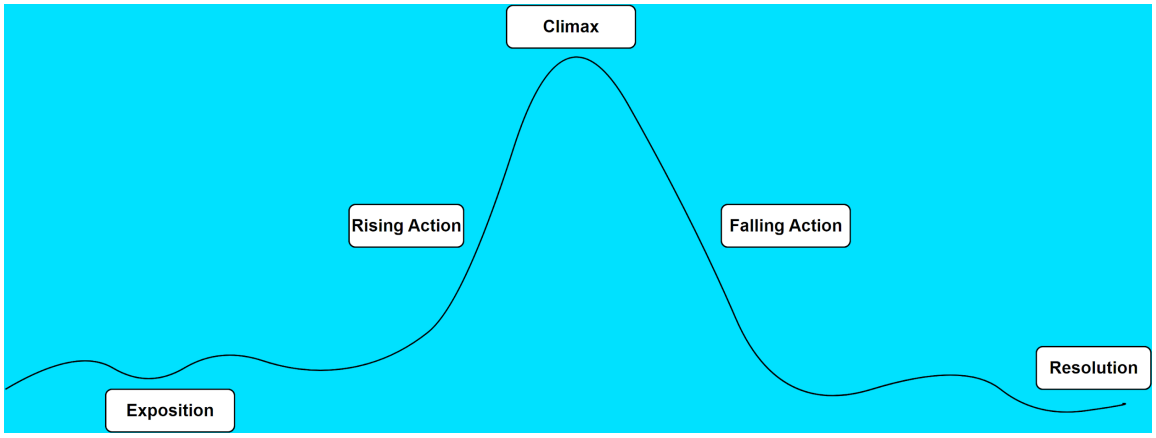


Figure 2. Narrative Curve

3.4. Modes of Operations

There will be a single mode of operation, which is the program itself written in Python being run on a compiler. The program will be used to generate the final model and report as needed by the team members.

3.5. Users

This project will not be available as an application for use by a user. The program will be only used by the team members to show to Los Alamos National Laboratories and Texas A&M. No training is necessary since it is only the team members who will be using the system. The institution that requested the project will receive a technical report describing the data, methods, and results with the final narrative structure model. The results and meaning behind them as well as the implications of the predictions can be explained and elaborated by the project team.

3.6. Support

A technical report will outline the methods, data, and conclusions reached by the project. This would include the framework of the scraper, NLP model, and gradient boosting model will be included so the methods used could be used for other applications. All program support will be handled in-house by the project team.

4. Scenario(s)

4.1. Institution

The institution asks for a public sentiment and subjectivity report. The project will generate a technical report based on the institution's parameters and topics that they want to explore. The institution can ask for an analysis of specific events or topics that would detail the public’s perception of the institution in regard to the specific topics or events. The parameters and topics can be preselected or defined such as certain events, time, news sources, topics, or people to determine how public perception has been affected and the different variables that went into that effect. This kind of project can look at trends and create predictions that could be used for risk management decision-making for the institution using the tool.

5. Analysis

5.1. Summary of Proposed Improvements

- News analytic tool using non-proprietary tools
- Seamless communication between the NLP and deep learning models
- Proposed system would be applicable to other institutions
- Local database for easy access

5.2. Disadvantages and Limitations

- The system would need a team to do the required training and application to a different institute which would take time to scrape data and train the model
- The Deep learning model would need continuous training and updates to keep the model relevant
- This project is not focused on creating a user application so a user interface would not be created to access the data code and methods used.

5.3. Alternatives

- Use premade news analytics software
    - Wouldn't have defined methods in extracting data, grading sentiment and predictions
    - Extra work would be needed to make the output of the software fit the needs of the user
- Model could answer a different question than what was defined
    - The model would have to be trained for this different question
    - The methods to create the model might not be viable for the different question

5.4. Impact

- General sentiment trends would give the institution an assessment on whether they maintain the public's trust
- Assigning numerical scores to abstract concepts can characterize the actual sentiment and subjectivity of an article
- Reports on public perception can directly change public opinion if released.

NEWS Analytics

Alexander Corpstein
James Nieberding

FUNCTIONAL SYSTEM REQUIREMENTS

REVISION – 2
03 December 2023

Functional System Requirements
FOR
News Analytics

Prepared by:

_____
Author Date

Approved by:

_____
Project Leader Date

_____
John Lusher, P.E. Date

_____
T/A Date

Change Record

Rev.	Date	Originator	Approvals	Description
1	9/28/2023	Tyler Shippy		First Draft
2	12/03/2023	Alexander Corpsein		Revised For Final Report
3	4/29/2024	James Nieberding		Revised For End of Project Report

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Purpose and Scope

This specification defines the technical requirements for the development items and support subsystems delivered to the client for the project. The purpose of this project is to develop methods to analyze the sentiment behind a certain institution. The project shall scrape news articles from the internet which will be stored on a database. It shall clean the documents of punctuation and capitalization as well as analyze the documents for sentiment and subjectivity along with other analytic scores. This data shall be sent to a gradient-boosting algorithm to perform narrative curve shaping as well as predict the shape of the curve moving into the future. This will all be combined to create a final technical report delivered to the sponsor outlining the results, methods, and data used.

The verification requirements for the project are contained in a separate Verification and Validation Plan.

1.2. Responsibility and Change Authority

As a team, we will be held responsible for our own subsystems and making sure we meet the requirements for each one. The team leader Alexander Corpstein is responsible for ensuring all system requirements are met, and the execution and validation plan is performed. Changes to requirements will only be done with the concurrence of all team members and the sponsor. Subsystems are owned as follows:

- Scraper, Article Cleaner, and Database Setup/Control: Alex Corpstein
- Sentiment and Subjectivity Analyzer, Topic Modeling Algorithm, and Narrative Curve Model: James Nieberding

2. Applicable and Reference Documents

2.1. Applicable Documents

The following documents, of the exact issue and revision shown, form a part of this specification to the extent specified herein:

No documents are used for specification requirements.

2.2. Reference Documents

The following documents are reference documents utilized in the development of this specification. These documents do not form a part of this specification and are not controlled by their reference herein.

Document Number	Revision/Release Date	Document Title
1	3.11.5	Python Library Reference
2	a320e5f6	Scrapy 2.10 Documentation
3	3.12.0	SQlite3 Documentation
4	v4.0	Selenium Documentation
5	v3.5.20357	SSHFS-Win
6	096047c5	dmlc XGBoost
7	3.8.1	NLTK Documentation
8	0.16.0.	Textblob Documentation
9	2.1.1	Pandas Documentation
10	2023-07-06	Tensorflow Documentation
11	4.3.0	Gensim Documentation
12	0.21.3	ScikitLearn Documentation
13	1.26	Numpy Documentation

2.3. Order of Precedence

In the event of a conflict between the text of this specification and an applicable document cited herein, the text of this specification takes precedence without any exceptions.

All specifications, standards, exhibits, drawings, or other documents that are invoked as “applicable” in this specification are incorporated as cited. All documents that are referred to within an applicable report are considered to be for guidance and information only, except ICDs that have their relevant documents considered to be incorporated as cited.

3. Requirements

This section defines the minimum requirements that the development item(s) must meet. The requirements and constraints that apply to performance, design, interoperability, reliability, etc., of the system, are covered.

3.1. System Definition

This project can be split into 4 main subsystems which can have smaller and more in-depth functionalities. The 4 subsystems are the Scraper, Database, NLP Sentiment/Subjectivity Analyzer, and finally the Narrative Curve Modeler. The Scraper is used to get the news article data that we need to analyze the sentiment of a certain institution as well as put it into a format that can be stored in our database. The Database is an SQL database hosted on the Olympus server that is responsible for storing all news article data coming from the scraper as well as the analyzed data coming from the analyzer. The Analyzer uses NLP cleaning and parsing techniques to score the article on its sentiment and subjectivity as well as other data points that can be measured. The Modeler is responsible for grouping the articles into topic buckets which can then be processed by a gradient-boosting model to create a narrative curve that fits the shape of the sentiment over time while also giving predictions based on its training.

This project is not made with the intent to create a user interface, so all methods for data collection and processing are determined by the creator. The institution chosen is determined when collecting the data with the scraper and being sent to the database.

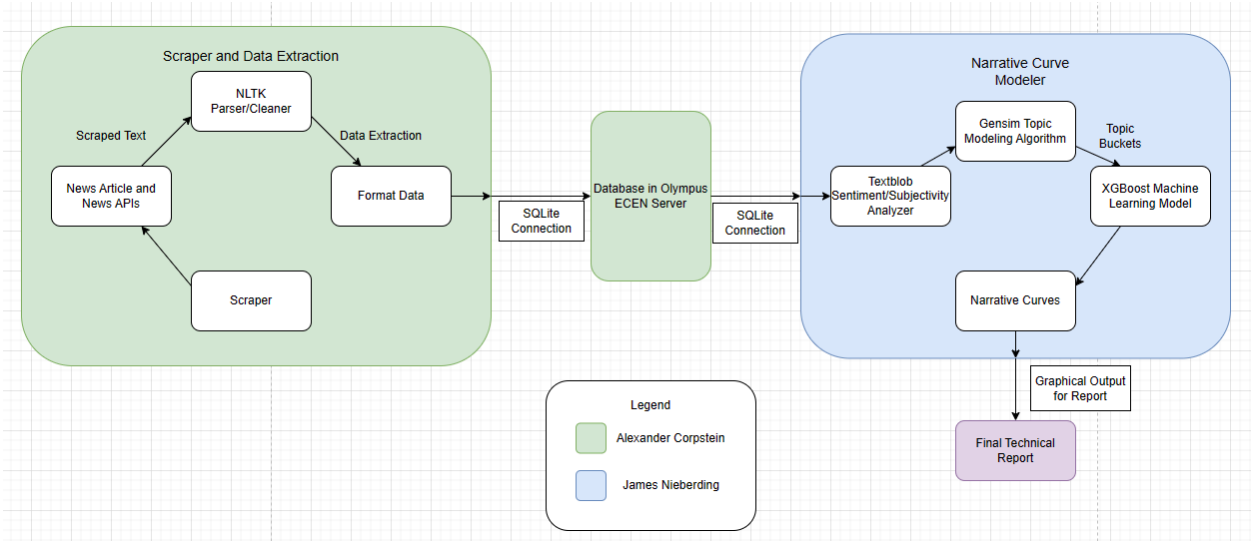


Figure 1. Block Diagram of System

3.1.1. Data Extraction and Database

The web scraper subsystem will be written using the Python packages Scrappy and Selenium. The scraper will be used to extract data from news articles on the internet and then format the data to be sent to the database.

A database will be used to store data for use by the NLP and Modeler subsystems. The database will be hosted on the TAMU ECEN Olympus server and created using SQlite3.

3.1.2. NLP Text Processing

Text from the database will be parsed for punctuation and cleaned in order to be used by the analyzer. Python's Natural Language Toolkit will be used to parse and clean the text, and Textblob will be used to analyze the text to produce sentiment and subjectivity scores. The emotion recognition component will analyze the text for each article and return the probabilities for each of the following emotions: sadness, joy, love, anger, fear, and surprise. This will be done using a predictive tensorflow machine learning model. This whole process will be done by importing a CSV, obtaining the data needed from it using pandas, processing that data, and finally outputting a new CSV containing all of the relevant information.

3.1.3. Predictive Narrative Curve Modeler

This subsystem is responsible for creating the output that will be used for the final technical report. It will first take the output data from the analyzer out of the database. Using Gensim, we will be able to analyze the words of the articles using a topic modeling algorithm. This then groups the articles into topic buckets which can be sent to the Narrative Curve Modeler. Based on trained data, predictive narrative curves over specific topics can be created and plotted. Articles can be clustered together where peaks in the curve are found to detail what events in the news led to the peak. This will then all be combined into the technical report for the sponsor at the end of the project to outline the methods used to create these curves and the implications of certain clusters of articles and the events associated with them.

3.2. Characteristics

3.2.1. Functional / Performance Requirements

3.2.1.1. Data Storage/Size

The Olympus server gives us 500GB of storage to use. The project should not exceed the storage however TAMU IT can give it more storage as needed.

> *Rationale: 500GB is the storage space given to the team and most likely will not be exceeded, but can be if required.*

3.2.1.2. Technical Report

The report shall have a wide range of narrative curves with respect to the institution chosen. The system should look to create at the very least 5 narrative curves to get as much coverage over the different topics and fields that Texas A&M is involved in.

Rationale: Having 5 curves would cover most topics and fields that TAMU is in the news for and will give comprehensive results of the university's public perception.

3.2.1.3. Predictive Model Accuracy

The Predictive Model system shall be able to predict the ups and downs of the curve of a topic that already exists.

Rationale: The error associated with the machine learning model decreases as the training dataset increases. If we are able to gather a wide range of articles over multiple known topics, we can monitor the training of the model to ensure it is moving in the right direction in regards to producing an accurate prediction. That way when applying the model to unknown topics we can have a higher level of confidence in its output.

3.2.2. Software Requirements

3.2.2.1. Software Installation

The project is all software and shall be contained either on the Olympus ECEN servers or our personal devices. The system will need to run on an operating system that can run the Python environment and packages. The program SSFHS-Win will be used to facilitate connection to the database hosted on the Olympus server.

Rationale: Running the program will require the ability to run the Python code. However, no specific operating system is required to do this. As well the code will be executed both inside the Olympus server and on local devices.

3.2.2.2. Python Environments and Packages

The following Python packages will be required to run the system: Gensim, Scrapy, Selenium, SGlite3, XGboost, NLTK, Pandas, Textblob, ScikitLearn, and Numpy.

Rationale: Conda will control the packages and environment, but not all the needed packages are default. As such specific packages will be required to be installed.

3.2.3. Communication Requirements

3.2.3.1. Olympus Server

All parts of the project shall connect to the Olympus server that hosts the database. Connecting to the SSH shell requires access granted by the TAMU IT department. The program SSHFS-Win will be used to provide a secure connection.

Rationale: The Olympus server is hosted by TAMU and will require TAMU permission. Access to the SSH shell and team directory will require software capable of performing the required actions.

3.2.3.2. Database Format

The database will be created using SQLite3, a package in the Python standard library. The database will store information that will contain the following items: headline, article, date posted, source website, neg, neu, pos, compound, sentiment, subjectivity, anger, joy, love, surprise, fear, and sadness

Rationale: Sqlite3 will create a database sizable for our needs to store the required information for the analyzer and modeler to use.

3.2.4. Failure Propagation

Faults and errors created by the system will be handled by the software code to handle all exceptions.

Rationale: As this is not an application any errors and faults shall be caught by the subsystem's exception handling.

3.2.5. Physical Characteristics

As a completely software-based project, the system will not have any physical requirements

Rationale: Physical characteristics of the laptop running the system have no effect on the operation of the system.

3.2.6. Electrical Characteristics

The system is completely software-based as such there are no electrical characteristic requirements.

Rationale: Electrical input, output, and power consumed will be based on the laptop running the system and have no effect on the operation of the system.

3.2.7. Environmental Requirements

As a software-based project it will have no environmental requirements.

Rationale: Power supplied to the laptop will be from the local power grid and not under the control of the team.

4. Support Requirements

The final output of the project will be a technical report describing the methods, data, and results of the system. To support the technical report all team members shall be available as necessary by the sponsor to answer all questions that arise from the report. There shall not be any support provided on how to use the system to any people outside the project members.

Any support required for the use of the Olympus server will require contacting the TAMU IT department through their help desk.

> *Rationale: The team will answer all the questions that the sponsor or any agency asks about the results of the system. However, the system is only designed to be operated by team members. The Olympus server is owned by TAMU and they will need to be contacted for any help or support that the project members can not provide.*

Appendix A: Acronyms and Abbreviations

NLP	Natural Language Processing
TAMU	Texas A&M University
IT	Information Technology
ECEN	Electrical and Computer Engineering
ICD	Interface Control Document
NLTK	Natural Language Toolkit
FSR	Functional System Requirements
SQL	Structured Query Language
CSV	Comma-Separated Values

Appendix B: Definition of Terms

XGBoost	A python library that provides gradient boosting algorithms to analyze data
Gensim	A topic modeling library that groups data points into clusters based on their relevance to each other and the topic they fit under
NLTK	A python library that provides a large set of tools for text data, including, but not limited to, tokenization, stemming, lemmatization, parsing, and sentiment analysis.
Textblob	A Python library that provides a simple API for natural language processing for things such as sentiment analysis and subjectivity.
Scrapy	A Python library that provides tools for web scraping.
Sqlite3	A Python library that provides a set of tools for creating and controlling a database.
Tensorflow	TensorFlow is an open-source machine learning framework developed by Google for building and training neural network models.

NEWS Analytics

Alexander Corpstein
James Nieberding

INTERFACE CONTROL DOCUMENT

REVISION – 2
03 December 2023

Interface Control Document
FOR
News Analytics

Prepared by:

_____
Author Date

Approved by:

_____
Project Leader Date

_____
John Lusher II, P.E. Date

_____
T/A Date

Change Record

Rev.	Date	Originator	Approvals	Description
1	09/28/23	James Nieberding		First Draft
2	12/03/2023	Alexander Corpsein		Revised For Final Report
3	4/29/2024	James Nieberding`		Revised For End of Project Report

Table of Contents

List of Tables

1. Overview

This document covers the interfaces between the project subsystems. All the subsystems will need to interface with the Olympus server and the database contained within.

2. References and Definitions

2.1. References

Document Number	Revision/Release Date	Document Title
1	3.11.5	Python Library Reference
2	a320e5f6	Scrapy 2.10 Documentation
3	3.12.0	SQlite3 Documentation
4	v4.0	Selenium Documentation
5	v3.5.20357	SSHFS-Win
6	096047c5	dmlc XGBoost
7	3.8.1	NLTK Documentation
8	0.16.0.	Textblob Documentation
9	2.1.1	Pandas Documentation
10	2023-07-06	Tensorflow Documentation
11	0.21.3	ScikitLearn Documentation
12	1.26	Numpy Documentation
13	4.3.0	Gensim Documentation

2.2. Definitions

XGBoost	A python library that provides gradient boosting algorithms to analyze data
Gensim	A topic modeling library that groups data points into clusters based on their relevance to each other and the topic they fit under
NLTK	A python library that provides a large set of tools for text data, including, but not limited to, tokenization, stemming, lemmatization, parsing, and sentiment analysis.
Textblob	A python library that provides a simple API for natural language processing for things such as sentiment analysis and subjectivity.

Scrapy	A python library that provides tools for web scraping.
Sqlite3	A python library that provides the set of tools for creating and controlling a database.
Tensorflow	TensorFlow is an open-source machine learning framework developed by Google for building and training neural network models.

3. Internet Data Extraction

In the scraper and data extraction subsystem, data will be extracted from news websites on the internet using the Python library Scrapy. The data to be extracted will be article headline, article content, date published. All data extracted will be formatted as required by the database using SQLite3 before being sent and stored in the database on the Olympus server.

4. Database and Olympus Interface

4.1. Database

The database will be created and controlled by SQLite3. All the subsystems will need to interface with the database using the commands provided by SQLite3. The individual subsystems will then convert and process the data as needed. The information to be stored will be the article headline, article content, date published, source website, sentiment, and subjectivity.

4.2. Olympus

The database will be hosted on the Olympus server in a directory for the team. Connection to the server is controlled by the TAMU IT department and access will be granted by them. The program SSHFS-Win will be used to provide a secure connection.

5. Gensim and XGBoost Interface

Gensim groups documents into topics based on how it is trained. It also has built-in functions that describe to the user what topics the different articles fit into. In order to create narrative curves over the top 10 topics within the database, the system will have an interface in the Modeler sub-system that would read the most frequent topics that Gensim

found and then use the document information function which lists the topics that the different articles fit into.

6. NLTK, Textblob, and Tensorflow Interface

NLTK (Natural Language Toolkit) in Python will clean and parse the raw text data from the Olympus database, which will then be used by Textblob to generate the sentiment and subjectivity scores. NLTK and Textblob will communicate in the same Jupyter Notebook terminal. Any changes made to the way NLTK parses and cleans the given text data will directly affect Textblob's output scores. Any changes made to the text cleaning from NLTK will also directly affect the model's output.

7. Physical Interface

As a software-based program there will be no physical interface between the project systems.

8. Thermal Interface

As a software-only project there will be no thermal interface between the subsystems.

9. Electrical Interface

As a software-only project all interfaces between the subsystems will be programmed based and there is no direct electrical interface between them.

NEWS Analytics

Alexander Corpstein

SubSystem Report: Scraper & Database

Change Record

Rev.	Date	Originator	Approvals	Description
1	12/03/23	Alexander Corpstein		First Draft
2	04/29/24	Alexander Corpstein		Revised for Final Report

SUBSYSTEM REPORT:

SCRAPER & DATABASE

FOR

News Analytics

TEAM <12>

APPROVED BY:

_____
Project Leader Date

_____
John Lusher II, P.E. Date

_____
T/A Date

Table of Contents

List of Tables

List of Figures

1. Subsystem Introduction

The scraper and database subsystem is made to scrape the desired data from news websites and store them into a database for access by the other subsystems. The data to be scraped is the article headline, the article content, the date published. This is stored in the database with entries for the sentiment and subjectivity models To do this the python libraries Selenium and Scarpy are used to build the scrapers and SQlite3 to make the database. The database is stored on Olympus, the Texas A&M ECEN server. To connect to the server the program SSHFS-WIn is used to mount the server as a local drive.

2. Software Requirements & Documentation

The subsystem is all software and shall be contained either on the Olympus ECEN servers or the team personal devices. The system will need to run on an operating system that can run the Python environment and packages.

The following Python packages will be required to run the system:, Scrapy, Selenium, SQlite3. The program SSHFS-Win is required to facilitate the connection to the Olympus server.

Table 1: References & Documentation

Document Number	Revision/Release Date	Document Title
1	3.11.5	Python Library Reference
2	a320e5f6	Scrapy 2.10 Documentation
3	3.12.0	SQlite3 Documentation
4	v4.0	Selenium Documentation
5	v3.5.20357	SSHFS-Win

3. Subsystem Functionality

The scraper program begins by using Selenium to open the target website and automatically search for or find the list of news articles that will be scraped. Then it is used to navigate through the list, recording the urls for each article and move to the

next page when required. This list of urls are then sent to a Scrapy function that interrogates the url to retrieve the headline, article, and date data. This is then sent to the function that will send the data to the database. In Figures 1 and 2 we can see an example webpage targeted by scrapy. By using the Xpath to the data that is to be collected, scrapy scrapes the text from the webpage. There are five scrapers made, one for the Texas Tribune, CNN, KBTX, Battalion, and Texas A&M Today.

Figure 1: Example Scrape Headline and Date

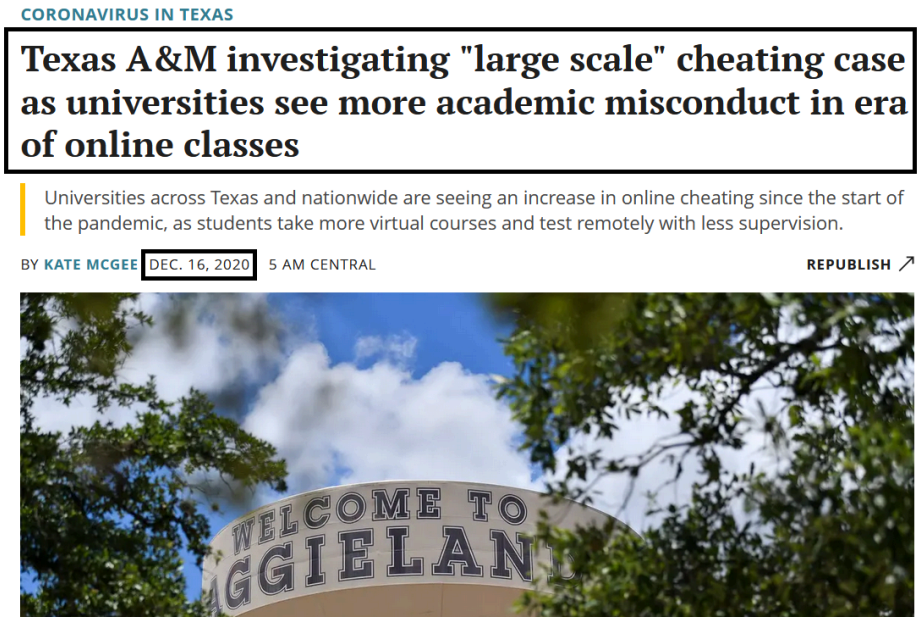
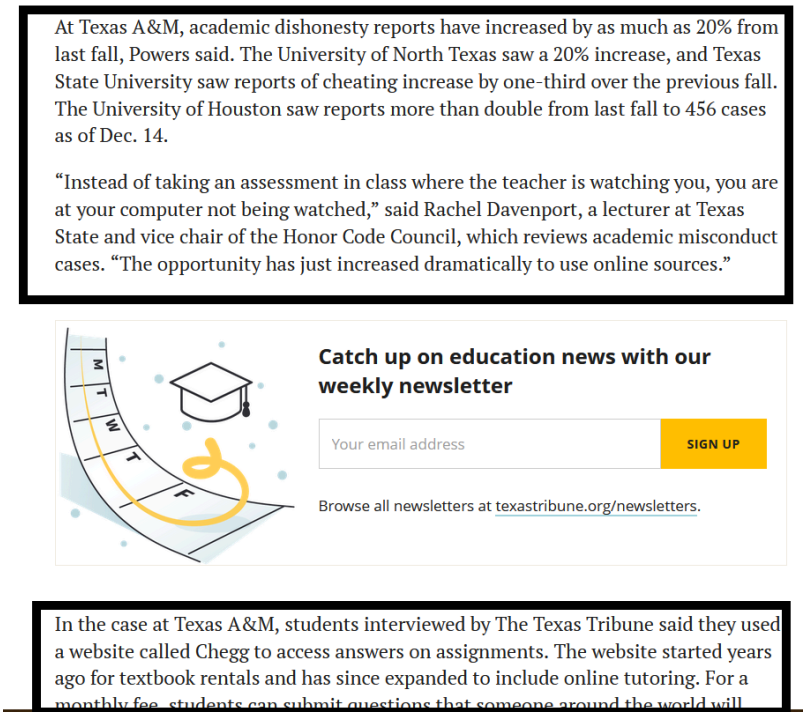


Figure 2: Example Scrape Article


At Texas A&M, academic dishonesty reports have increased by as much as 20% from last fall, Powers said. The University of North Texas saw a 20% increase, and Texas State University saw reports of cheating increase by one-third over the previous fall. The University of Houston saw reports more than double from last fall to 456 cases as of Dec. 14.

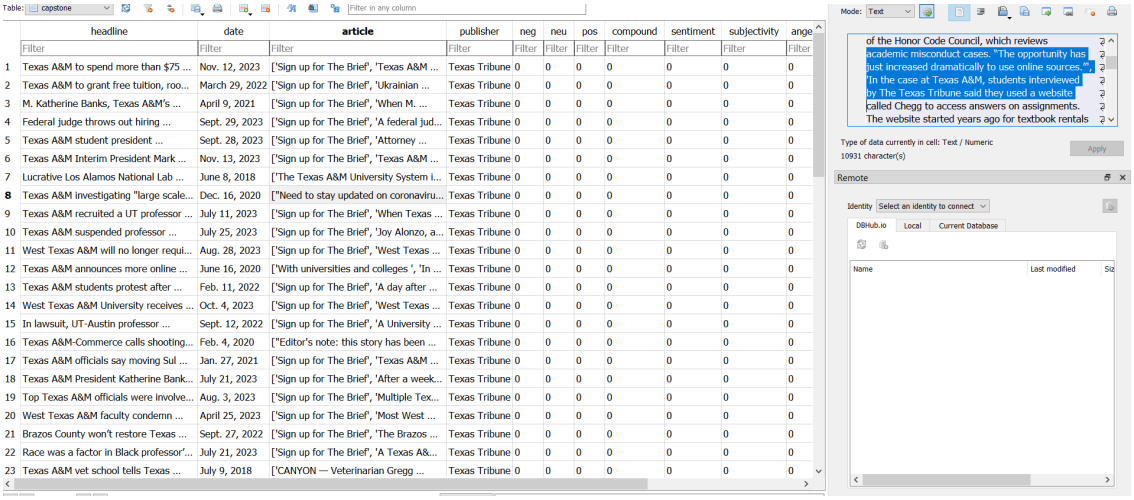
"Instead of taking an assessment in class where the teacher is watching you, you are at your computer not being watched," said Rachel Davenport, a lecturer at Texas State and vice chair of the Honor Code Council, which reviews academic misconduct cases. "The opportunity has just increased dramatically to use online sources."

In the case at Texas A&M, students interviewed by The Texas Tribune said they used a website called Chegg to access answers on assignments. The website started years ago for textbook rentals and has since expanded to include online tutoring. For a monthly fee, students can submit questions that someone around the world will

4. Database Function

The database is created using SQlite3, a package included in python by default. The table is created and updated through SQL language. The data is taken from the scraper function, then a connection is made to the database storage location on the Olympus server using SSFHS-Win. The data is formatted for insertion and placed in the database with all entries pertaining to the various models set to zero. The database stores the headline, article content, date published, publisher, neg, neu, pos, compound, sentiment, subjectivity, anger, joy, love, surprise, fear, and sadness. In Figure 3 we see an example of how the database looks using the DB Browser software. This software is not required for program use, but used for ease of testing and viewing the database. In the figure we can see how the database stores each entry for use by the other subsystem.

Figure 3: Example Database Viewed with DB Browser

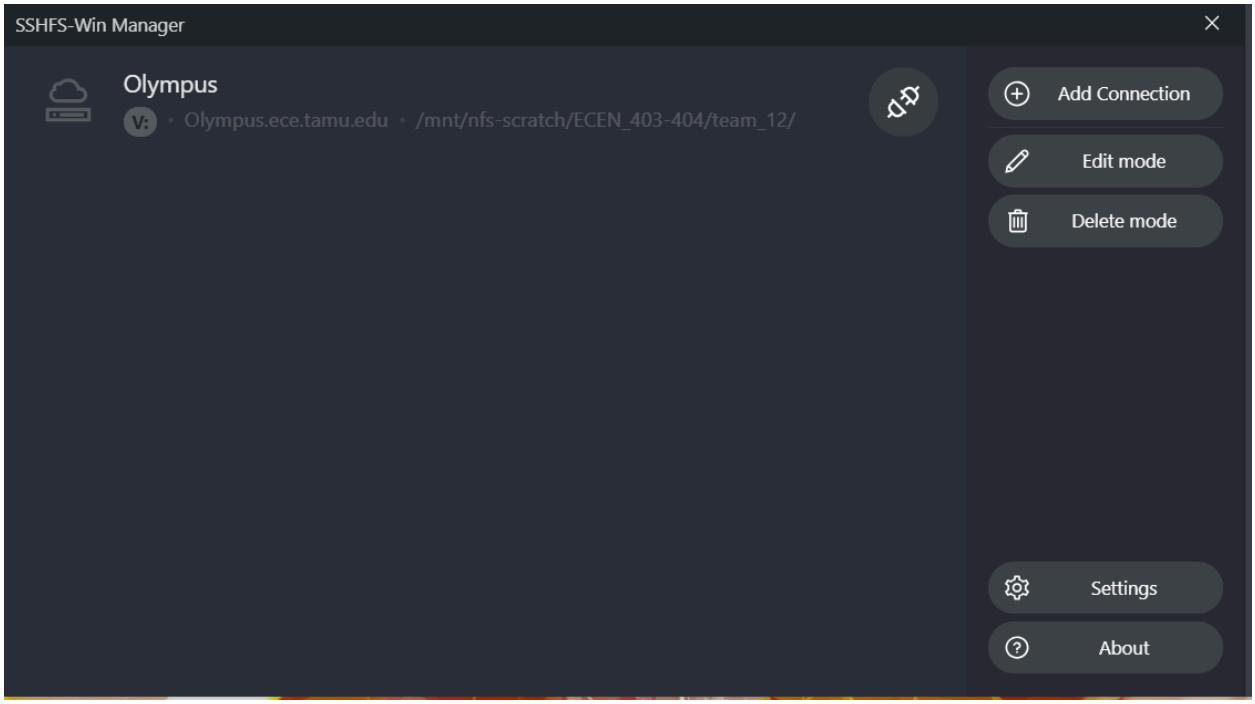


	headline	date	article	publisher	neg	neu	pos	compound	sentiment	subjectivity	ange
1	Texas A&M to spend more than $75 ...	Nov. 12, 2023	['Sign up for The Brief', 'Texas A&M ...	Texas Tribune	0	0	0	0	0	0	0
2	Texas A&M to grant free tuition, roo...	March 29, 2022	['Sign up for The Brief', 'Ukrainian ...	Texas Tribune	0	0	0	0	0	0	0
3	M. Katherine Banks, Texas A&M's ...	April 9, 2021	['Sign up for The Brief', 'When M. ...	Texas Tribune	0	0	0	0	0	0	0
4	Federal judge throws out hiring ...	Sept. 29, 2023	['Sign up for The Brief', 'A federal jud...	Texas Tribune	0	0	0	0	0	0	0
5	Texas A&M student president ...	Sept. 28, 2023	['Sign up for The Brief', 'Attorney ...	Texas Tribune	0	0	0	0	0	0	0
6	Texas A&M Interim President Mark ...	Nov. 13, 2023	['Sign up for The Brief', 'Texas A&M ...	Texas Tribune	0	0	0	0	0	0	0
7	Lucrative Los Alamos National Lab ...	June 8, 2018	['The Texas A&M University System i...	Texas Tribune	0	0	0	0	0	0	0
8	Texas A&M investigating "large scale...	Dec. 16, 2020	["Need to stay updated on coronaviru...	Texas Tribune	0	0	0	0	0	0	0
9	Texas A&M recruited a UT professor ...	July 11, 2023	['Sign up for The Brief', 'When Texas ...	Texas Tribune	0	0	0	0	0	0	0
10	Texas A&M suspended professor ...	July 25, 2023	['Sign up for The Brief', 'Joy Alonzo, a...	Texas Tribune	0	0	0	0	0	0	0
11	West Texas A&M will no longer requi...	Aug. 28, 2023	['Sign up for The Brief', 'West Texas ...	Texas Tribune	0	0	0	0	0	0	0
12	Texas A&M announces more online ...	June 16, 2020	['With universities and colleges ', 'In ...	Texas Tribune	0	0	0	0	0	0	0
13	Texas A&M students protest after ...	Feb. 11, 2022	['Sign up for The Brief', 'A day after ...	Texas Tribune	0	0	0	0	0	0	0
14	West Texas A&M University receives ...	Oct. 4, 2023	['Sign up for The Brief', 'West Texas ...	Texas Tribune	0	0	0	0	0	0	0
15	In lawsuit, UT-Austin professor ...	Sept. 12, 2022	['Sign up for The Brief', 'A University ...	Texas Tribune	0	0	0	0	0	0	0
16	Texas A&M-Commerce calls shooting...	Feb. 4, 2020	["Editor's note: this story has been ...	Texas Tribune	0	0	0	0	0	0	0
17	Texas A&M officials say moving Sul ...	Jan. 27, 2021	['Sign up for The Brief', 'Texas A&M ...	Texas Tribune	0	0	0	0	0	0	0
18	Texas A&M President Katherine Bank...	July 21, 2023	['Sign up for The Brief', 'After a week...	Texas Tribune	0	0	0	0	0	0	0
19	Top Texas A&M officials were involve...	Aug. 3, 2023	['Sign up for The Brief', 'Multiple Tex...	Texas Tribune	0	0	0	0	0	0	0
20	West Texas A&M faculty condemn ...	April 25, 2023	['Sign up for The Brief', 'Most West ...	Texas Tribune	0	0	0	0	0	0	0
21	Brazos County won't restore Texas ...	Sept. 27, 2022	['Sign up for The Brief', 'The Brazos ...	Texas Tribune	0	0	0	0	0	0	0
22	Race was a factor in Black professor'...	July 21, 2023	['Sign up for The Brief', 'A Texas A&...	Texas Tribune	0	0	0	0	0	0	0
23	Texas A&M vet school tells Texas ...	July 9, 2018	['CANYON — Veterinarian Gregg ...	Texas Tribune	0	0	0	0	0	0	0

5. Scraper Database Interface

SSHFS-Win is used to mount to the Olympus server as a local drive. By doing this we are able to make use of SQlite3, a local database package on a remote server. SSHFS-Win makes this connection using Cygwin for the POSIX environment and WinFsp for the FUSE functionality. In Figure 4 we can see Olympus mounted as local drive V:. This allows for the file path of the database to be found by SQlite3 and data to be sent from the program to the server database.

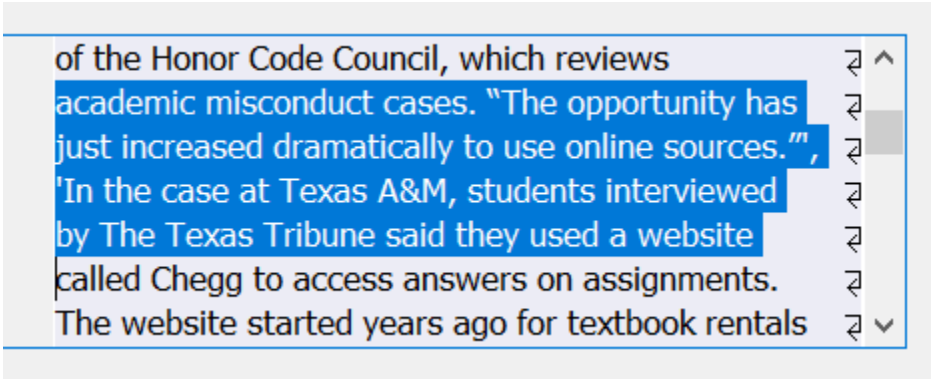
Figure 4: Olympus Server Connection via SSHFS-WIn



6. Validation Testing & Results

To validate the scraper and database subsystem an examination of the data is required. By looking at entries into database use DB Browser and grabbing the output as a csv file it was possible to examine how well the subsystem worked. For example by looking at Figures 1 and 5 we can see how the text of the news article was copied into the database in its entirety while missing the image that broke up the paragraphs.

Figure 5: Example Article Entry in Database



This is done for each scraper to verify that during the test scrapes the data in the database matches the article it came from. This also reveals bugs such as with the KBTX scraper. In Figure 6 we can see the test database for the KBTX data and can see that while the method employed by the scraper did pick up article content it was

not the actual article as none existed for sports information. This is a bug discovered by the process of manually checking the content being scrapped as testing occurs.

Figure 6: Example Entry from KBTX Database

56	Texas A&M vs. Cal November 25 Women's Basketball Tickets & Start Time	Nov. 25, 2023 at 12:07 AM CST	['© 2023 Data Skrive. All rights ...	KBTX

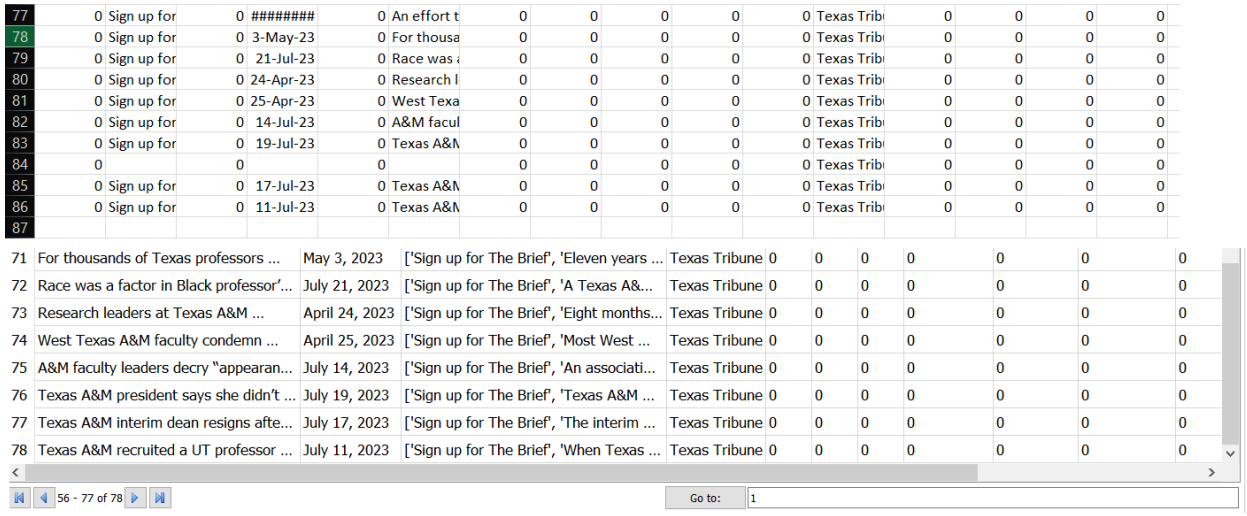
The database automatically filters out incomplete entries into the database. As such another way to test the subsystem is to examine what webpages the scraper missed. In Figure 7 we can see an example of the CSV output which does not take the data not sent to the database, but the output of the console and converts it to a csv file. From the figure we can see a miss from one of the entries for the article content. From this the program is updated to account for the difference in web pages inside the news site. However not every scraped url is an article. As seen in the KBTX example some urls will not contain articles, be videos, links to other webpages, or anything else that is not an actual article from the website.

Figure 7: Example CSV Output

anger	article	compound	date	fear	headline	joy	love	neg	neu	pos	publisher	sadness	sentiment	subjectivit	suprise
0	Â ,TheÂ ,M	0	########	0	Chromic D	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,An imm	0	########	0	â€˜Patient	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,The new	0	########	0	David Parr	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,Texas A	0	14-Jun-23	0	Texas A&M	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,As a Lat	0	7-Jun-23	0	â€˜Across	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,The ,As	0	5-Sep-23	0	School Of	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,Thereâ€	0	2-Aug-23	0	Summer R	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,Texas A	0	########	0	Texas A&M	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,As the tl	0	########	0	Climate Sc	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,Dance c	0	########	0	Artists Fro	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,The Rud	0	11-Apr-23	0	Venture To	0	0	0	0	0	Texas A&M	0	0	0	0
0		0	26-Apr-23	0	Visualizati	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,Looking	0	8-May-23	0	LIVE Lab E	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,KAMU-T	0	########	0	Aggie-Proc	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,Texas A	0	18-Jan-23	0	Texas A&M	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,A groggy	0	27-Jan-23	0	Texas A&M	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,At a hun	0	########	0	Aggie Arch	0	0	0	0	0	Texas A&M	0	0	0	0
0	Â ,A Texas	0	########	0	Exhibition	0	0	0	0	0	Texas A&M	0	0	0	0
0	This summ	0	########	0	Texas A&M	0	0	0	0	0	Texas A&M	0	0	0	0
0	Funded by	0	8-Sep-22	0	Six Animat	0	0	0	0	0	Texas A&M	0	0	0	0

By examining the differences between the csv output and the total number of entries into the test databases it is also possible to get an accuracy of the scraper as well as see how many urls were missed when the list of them was being gathered. In Figure 8 it is seen that the csv recorded 86 entries, but 78 were sent to the database. This is out of an attempt to scrape 100 articles.

Figure 8: Example CSV vs Database Entry



77	0	Sign up for	0	########	0	An effort t	0	0	0	0	0	Texas Trib	0	0	0	0
78	0	Sign up for	0	3-May-23	0	For thousa	0	0	0	0	0	Texas Trib	0	0	0	0
79	0	Sign up for	0	21-Jul-23	0	Race was a	0	0	0	0	0	Texas Trib	0	0	0	0
80	0	Sign up for	0	24-Apr-23	0	Research l	0	0	0	0	0	Texas Trib	0	0	0	0
81	0	Sign up for	0	25-Apr-23	0	West Texa	0	0	0	0	0	Texas Trib	0	0	0	0
82	0	Sign up for	0	14-Jul-23	0	A&M facul	0	0	0	0	0	Texas Trib	0	0	0	0
83	0	Sign up for	0	19-Jul-23	0	Texas A&M	0	0	0	0	0	Texas Trib	0	0	0	0
84	0		0		0		0	0	0	0	0	Texas Trib	0	0	0	0
85	0	Sign up for	0	17-Jul-23	0	Texas A&M	0	0	0	0	0	Texas Trib	0	0	0	0
86	0	Sign up for	0	11-Jul-23	0	Texas A&M	0	0	0	0	0	Texas Trib	0	0	0	0
87																

71	For thousands of Texas professors ...	May 3, 2023	['Sign up for The Brief', 'Eleven years ...	Texas Tribune	0	0	0	0	0	0	0
72	Race was a factor in Black professor'...	July 21, 2023	['Sign up for The Brief', 'A Texas A&...	Texas Tribune	0	0	0	0	0	0	0
73	Research leaders at Texas A&M ...	April 24, 2023	['Sign up for The Brief', 'Eight months...	Texas Tribune	0	0	0	0	0	0	0
74	West Texas A&M faculty condemn ...	April 25, 2023	['Sign up for The Brief', 'Most West ...	Texas Tribune	0	0	0	0	0	0	0
75	A&M faculty leaders decry "appearan...	July 14, 2023	['Sign up for The Brief', 'An associati...	Texas Tribune	0	0	0	0	0	0	0
76	Texas A&M president says she didn't ...	July 19, 2023	['Sign up for The Brief', 'Texas A&M ...	Texas Tribune	0	0	0	0	0	0	0
77	Texas A&M interim dean resigns afte...	July 17, 2023	['Sign up for The Brief', 'The interim ...	Texas Tribune	0	0	0	0	0	0	0
78	Texas A&M recruited a UT professor ...	July 11, 2023	['Sign up for The Brief', 'When Texas ...	Texas Tribune	0	0	0	0	0	0	0

56 - 77 of 78 Go to: 1

By examining the difference between these values Table 2 is developed. From this table it is possible to get an idea of the accuracy of the scrapers. This table does not account for the problem stated before regards to a non article content that is targeted by the scraper. It also does not account for any missed urls due to connection and internet problems. Urls that are actually missed by the scraper are accounted for in an error text file that is generated when the program is run. This is seen in Figure 9 and it is noted that the file also contains entries for missed scraped data.

Table 2: Accuracy of Five Scrapers

Website	# of Articles Attempted to Scrape	# of Entries in CSV	# of Entries in Database
Texas Tribune	100	86	78
CNN	100	83	52
KBTX	100	91	65
Battalion	100	73	69
Today	100	100	93

Figure 9: Error Text File

```
Batt
Miss article https://www.thebatt.com/sports/texas-a-m-falls-to-texas-southern-at-home/article_45dd187e-0bd2-11e9-a809-bb05080e80dc.html
Miss article https://www.thebatt.com/sports/texas-a-m-dominates-texas-southern/article_e997545e-3d6e-11e7-8704-b70894df400b.html
Miss article https://www.thebatt.com/sports/texas-a-m-falls-to-texas-southern-at-home/article_45dd187e-0bd2-11e9-a809-bb05080e80dc.html
Miss article https://www.thebatt.com/sports/texas-a-m-dominates-texas-southern/article_e997545e-3d6e-11e7-8704-b70894df400b.html
missed url 1 page 3
missed url 1 page 3
missed url 4 page 3
missed url 5 page 3
missed url 6 page 3
missed url 7 page 3
missed url 8 page 3
missed url 9 page 5
missed url 9 page 6
missed url 9 page 7
missed url 9 page 8
Miss article https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=150
Miss headline https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=150
Miss date https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=150
Miss article https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=125
Miss headline https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=125
Miss date https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=125
Miss article https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=175
Miss headline https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=175
Miss date https://www.thebatt.com/search/?f=html&q=Texas+A%26M&sd=desc&l=25&t=article&nsa=eedition&app%5B0%5D=editorial&o=175
kbtx
missed url 12 page 3
missed url 13 page 3
missed url 14 page 3
missed url 15 page 3
Miss article https://www.kbtx.com/video/2023/11/28/texas-ams-inaugural-winter-wanderland-lights-up-aggie-park/
Miss headline https://www.kbtx.com/video/2023/11/28/texas-ams-inaugural-winter-wanderland-lights-up-aggie-park/
Miss date https://www.kbtx.com/video/2023/11/28/texas-ams-inaugural-winter-wanderland-lights-up-aggie-park/
Miss date https://www.kbtx.com/sports/betting/2024/03/01/texas-a-m-cc-womens-college-basketball-march-madness-odds/
Miss date https://www.kbtx.com/sports/betting/2023/07/01/texas-a-m-aggies-2023-college-football-schedule-how-to-watch/
Miss article https://www.kbtx.com/video/2023/11/27/mike-elko-texas-am/
Miss headline https://www.kbtx.com/video/2023/11/27/mike-elko-texas-am/
Miss date https://www.kbtx.com/video/2023/11/27/mike-elko-texas-am/
```

This error file is not only useful for finding and correcting missed urls and data entries when possible, but will be used when performing large scale scrapes to know where to continue from if a problem occurs.

With these tests the validation plan was carried out. The data scraped was manually verified to be correct and from the target urls. The database was created and successfully stores the data formatted correctly for future use by the other subsystems. The data in the database is verified from the scraper output and can be reached for when needed by integration. This was done while the database was on the Olympus server verifying that connection and access.

7. Conclusion

The scraper and database subsystem is functional and ready for integration with the other subsystems. There are additional bugs and fixes that can be made to improve the accuracy of the scrapers, but that will not affect the integration of the other subsystems to the database. The scrapers work and will be able to be used to fill the database to be used by the project with sufficient entries for use by the models of the other subsystems.

8. Future Planning

With the scrapers tested and verified the next step will be filling a database with a large amount of entries for use by the other subsystems. While this is in progress bugs and problems noted in the testing will be fixed and improved where possible. The database is the point of integration for the other subsystem as such, work will be performed to aid the other subsystems with the needed integration. The experience gained while creating the five scrapers tested can be used to expand the number of scrapers providing even more data for use by the other subsystems.

NEWS Analytics

Alexander Corpstein
James Nieberding

SubSystem Report: Sentiment, Subjectivity, & Text Cleaner

Change Record

Rev.	Date	Originator	Approvals	Description
1	12/03/23	Tyler Shippy		First Draft
2	4/29/24	Alexander Corpstein		Revised for Final Report

SUBSYSTEM REPORT:

SENTIMENT, SUBJECTIVITY, & TEXT CLEANER

FOR

News Analytics

TEAM <12>

APPROVED BY:

_____

Project Leader Date

_____

John Lusher II, P.E. Date

_____

T/A Date

Table of Contents

List of Tables

List of Figures

2.1 Subsystem Introduction

The Sentiment, Subjectivity, and Text Cleaner Subsystem is designed to extract meaningful information from news articles utilizing NLP and Machine Learning techniques. The subsystem is broken up into two components: The Sentiment/Subjectivity component and the Text Cleaner component. This system was validated on various datasets to confirm its functionality and consistency.

2.2 Software Requirements and Documentation

The project is all software and shall be contained either on the Olympus ECEN servers or our personal devices. The system will need to run on an operating system that can run the Python environment and packages.
The following Python packages will be required to run the system: NLTK, TextBlob, Pandas, Tensorflow, ScikitLearn, and Numpy.

Table 1: Reference Documentation

Document Number	Revision/Release Date	Document Title
1	3.11.5	Python Library Reference
2	3.8.1	NLTK Documentation
3	0.16.0.	TextBlob Documentation
4	2.1.3	Pandas Documentation
5	2.14.0	TensorFlow Documentation
6	0.21.3	ScikitLearn Documentation
7	1.26	Numpy Documentation

2.3 Subsystem Details

2.2.1. Text Processing, NLTK, and TextBlob

This component of the subsystem involved the utilization of Python's NLTK (Natural Language Toolkit) and Textblob to analyze article data. The high-level process entailed reading a CSV, conducting an analysis of the text within, and generating a new CSV incorporating the relevant scores. For each article entry, the code standardizes the text to lowercase, substitutes ampersands with "and," eliminates non-alphabetic characters, excludes stopwords, performs lemmatization, and calculates sentiment scores using the NLTK library and TextBlob.

2.4 Subsystem Validation

2.4.1. Text Processing, NLTK, and TextBlob

Figure 1: First 20 entries of The Tribune news articles before code is run

Index	headline	article	date	publisher	sentimen	subjectivi	emotion
1	M. Katherine Banks, Texas A&Mâ€™s incoming president, says sheâ€™s planning for a	Sign up fo	9-Apr-21	Texas Trib	0	0	joy
2	Texas A&M investigating "large scale" cheating case as universities see more academi	Need to s	Dec. 16, 2(	Texas Trib	0	0	fear
3	Texas A&M to grant free tuition, room and board to its Ukrainian students	Sign up fo	########	Texas Trib	0	0	joy
4	Texas A&M announces more online offerings, classroom capacity caps and daily cleani	With univ	16-Jun-20	Texas Trib	0	0	joy
5	Lucrative Los Alamos National Lab contract awarded to team that includes Texas A&M	The Texas	8-Jun-18	Texas Trib	0	0	joy
6	Federal judge throws out hiring discrimination lawsuit against Texas A&M	Sign up fo	Sept. 29, 2	Texas Trib	0	0	sadness
7	Texas A&M President Katherine Banks resigns amid fallout from failed hiring of journa	Sign up fo	21-Jul-23	Texas Trib	0	0	sadness
8	Texas A&M suspended professor accused of criticizing Lt. Gov. Dan Patrick in lecture	Sign up fo	25-Jul-23	Texas Trib	0	0	sadness
9	Texas A&M officials say moving Sul Ross statue is no longer an option, but students say	Sign up fo	Jan. 27, 20	Texas Trib	0	0	surprise
10	Texas A&M student president impeached, removed from office	Sign up fo	Sept. 28, 2	Texas Trib	0	0	sadness
11	TribCast: Turmoil at Texas A&M	In this we	Aug. 4, 20	Texas Trib	0	0	anger
12	West Texas A&M will no longer require students to pay for textbooks starting next fall	Sign up fo	Aug. 28, 2(	Texas Trib	0	0	joy
13	Brazos County wonâ€™t restore Texas A&M early-voting location despite studentsâ€™	Sign up fo	Sept. 27, 2	Texas Trib	0	0	anger
14	Accountability U.	Texas Trib	Sept. 9, 2(	Texas Trib	0	0	joy
15	Texas A&M recruited a UT professor to revive its journalism program, then backtracke	Sign up fo	11-Jul-23	Texas Trib	0	0	sadness
16	Texas A&M faculty leaders say President Kathy Banks is leaving them out of major deci	Sign up fo	Aug. 10, 2(	Texas Trib	0	0	surprise
17	Texas A&M students protest after president ends print publication of 129-year-old Bat	Sign up fo	Feb. 11, 2(	Texas Trib	0	0	anger
18	Texas A&M and University of Texas systems expect to reopen in the fall, and A&M says	Editor's nc	########	Texas Trib	0	0	joy
19	Texas A&M System employees were asked to teach office pet birds to say â€œhowdy.	COLLEGE S	Oct. 14, 2(	Texas Trib	0	0	sadness
20	Interim Texas A&M University president sets new tone for reforms planned under his	Sign up fo	Oct. 4, 202	Texas Trib	0	0	joy

Figure 2: First 20 entries of The Tribune news articles after code is run

Index	headline	article	date	publisher	sentimen	subjectivi	emotion	neg	neu	pos	compoun	Sentimen	Subjectivity
1	katherine	sign kathe	9-Apr-21	Texas Trib	0	0	joy	0.065	0.774	0.162	0.9976	0.087037	0.379584
2	texas inve	need stay	Dec. 16, 2(	Texas Trib	0	0	fear	0.109	0.74	0.15	0.9766	0.071396	0.35415
3	texas gran	sign stude	########	Texas Trib	0	0	joy	0.116	0.677	0.207	0.9022	-0.0254	0.495896
4	texas ann	university	16-Jun-20	Texas Trib	0	0	joy	0.025	0.904	0.071	0.7096	0.026317	0.305447
5	lucrative l	texas univ	8-Jun-18	Texas Trib	0	0	joy	0.027	0.765	0.208	0.9897	0.057411	0.348979
6	federal ju	sign feder	Sept. 29, 2	Texas Trib	0	0	sadness	0.084	0.813	0.103	0.4767	0.046169	0.261944
7	texas pres	sign week	21-Jul-23	Texas Trib	0	0	sadness	0.089	0.783	0.128	0.9829	0.095774	0.432818
8	texas susp	sign respe	25-Jul-23	Texas Trib	0	0	sadness	0.11	0.777	0.113	-0.4019	0.02669	0.388441
9	texas offi	sign leade	Jan. 27, 20	Texas Trib	0	0	surprise	0.053	0.853	0.094	0.9359	0.007014	0.415225
10	texas stud	sign gener	Sept. 28, 2	Texas Trib	0	0	sadness	0.022	0.829	0.149	0.9738	0.113333	0.295556
11	turmoil te	discus two	Aug. 4, 20	Texas Trib	0	0	anger	0	0.756	0.244	0.6369	0.2	0.5
12	west texa	sign texas	Aug. 28, 2(	Texas Trib	0	0	joy	0.019	0.809	0.172	0.9978	0.105718	0.413819
13	brazos cou	sign brazo	Sept. 27, 2	Texas Trib	0	0	anger	0.055	0.808	0.137	0.9896	0.065019	0.315619
14	accountab	texas tribu	Sept. 9, 2(	Texas Trib	0	0	joy	0	1	0	0	-0.03333	0.366667
15	texas recr	sign texas	11-Jul-23	Texas Trib	0	0	sadness	0.046	0.769	0.185	0.9987	0.102348	0.383102
16	texas facu	sign leade	Aug. 10, 2(	Texas Trib	0	0	surprise	0.076	0.82	0.103	0.7964	0.06282	0.322602
17	texas stud	sign day te	Feb. 11, 2(	Texas Trib	0	0	anger	0.056	0.799	0.145	0.9836	0.061758	0.345241
18	texas univ	story upda	########	Texas Trib	0	0	joy	0.029	0.814	0.157	0.9962	0.042138	0.387304
19	texas syst	college st	Oct. 14, 2(	Texas Trib	0	0	sadness	0.02	0.748	0.232	0.9746	0.116026	0.49359
20	interim te	sign interi	Oct. 4, 202	Texas Trib	0	0	joy	0.104	0.797	0.099	-0.7259	0.063285	0.424796

Figure 3: First entry of The Tribune scraped article text

Sign up for The Brief,When M. Katherine Banks became the new engineering school dean at Texas A&M University in 2012, an associate dean drove her to the Zachry building, the heart of the universityâ€™s engineering program in College Station.,He pulled into the basement-level garage and honked the horn twice â€" a preventive measure to scare off skunks that infested the aging facility.,Above them, the building reeked of mold. Engineering students not-so-lovingly dubbed it the â€œZachry Smell.â€ Inside the building's lecture hall, a once-impressive revolving stage, was broken.,Ten years later, the skunks are gone. Banks raised $76 million in private donations that helped build a state-of-the-art engineering building that boasts multiple high-tech labs and collaborative meeting spaces.,She commissioned 10 original art pieces, all inspired by science, technology and engineering, for the half a million square foot building that now enrolls more than 21,000 students â€" nearly twice the number of students from when she started. Many leaders within Texas A&M point to the center as a physical testament to her success running the college, which now enrolls nearly a third of all of Texas A&Mâ€™s student body.,â€œThe house that Banks built,â€ an assistant vice chancellor joked last week during a tour of the facility, causing Banks to cringe.,â€œI donâ€™t know where that [phrase] came from," she said modestly.,Banks is not typically one to seek out attention, but starting June 1 sheâ€™ll enter the spotlight as the 26th president of Texasâ€™ biggest university, the second woman to ever run the flagship campus.,She takes over as Texas A&M faces multiple challenges, including how to navigate a return to â€œnormalâ€ campus operations after the COVID-19 pandemic upended learning and campus life. Sheâ€™ll also take the reins of a diversity and inclusion plan meant to increase students and faculty of color after a year where conversations about racial injustice and inequality on campus took center stage.,The Board of Regents last week approved her as the next president of Texas A&M, with a $925,000 annual salary over the next five years. Banks ended up negotiating her salary down $350,000 less than she was initially offered due to the current economic situation brought by the pandemic.,â€œIf she can do for the rest of the university what she did for engineering, it's gonna be a hell of a show to watchâ€ said System Chancellor John Sharp, who said he had tried to convince Banks to apply for president six years ago, but she wasnâ€™t ready to leave the engineering school. â€œShe's a visionary and an executorâ€¦[but] at the core of it all, itâ€™s giving opportunities to the students that are here. I think that's what makes her tick.â€ ,While Texas A&M leaders said Banks has the track record of a proven leader, in an

Figure 4: First entry snippet of tribune articles, after the code is run

sign katherine bank became new engineering school dean texas university associate dean drove zachry heart engineering program college pulled garage honked horn twice preventive measure scare skunk infested aging building reeked engineering student dubbed inside lecture revolving year skunk bank raised million private donation helped build engineering building boast multiple lab collaborative meeting commissioned original art inspired technology half million square foot building enrolls student nearly twice number student many leader within texas point center physical testament success running enrolls nearly third texas student house bank assistant vice chancellor joked last week tour causing bank know came said typically one seek starting june enter spotlight president biggest second woman ever run flagship take texas face multiple including navigate return campus operation pandemic upended learning campus also take rein diversity inclusion plan meant increase student faculty color year conversation racial injustice inequality campus took center board regent last week approved next president texas annual salary next five bank ended negotiating salary le initially offered due current economic situation brought rest university gonna hell show said system chancellor john said tried convince bank apply president six year ready leave engineering visionary core giving opportunity student think make texas leader said bank track record proven interview texas bank tread cautiously reluctant even outline vision broader university discus goal declined weigh heated issue certainly address coming like debate lawrence sullivan ross even stopped short definitively choosing favorite piece art zachry building believe would presumptive outline vision without first meeting listening many constituent stakeholder opportunity bank wrote response follow said plan embark listening tour across campus making concrete plan setting admits learning curve ahead come running aspect including student affair bank clear want fall semester texas something return hopeful full stadium walk around hopeful social bank said emergency adjust right perhaps aggie planning fully operational campus normal said student learn best classroom interacting faculty may information need graduate cautioned simply returning college life anticipates much fall include

This procedure as shown in Figures 1-4 is iterated multiple times across datasets utilized throughout the project. The above includes a snapshot of a CSV before and after the execution of the code above, along with the raw and cleaned text of the first article in the dataset. In Figure 4, the ‘neg’, ‘neu’, ‘pos’, and ‘compound’ columns are the negative, neutral, positive, and overall scores from NLTK for the Sentiment in the article text. The ‘Sentiment’ and ‘Subjectivity’ columns are the Sentiment and Subjectivity scores generated

from TextBlob. The headlines as well as the article content are cleaned and replaced into a new CSV.

2.5. Conclusion

In conclusion, the Sentiment, Subjectivity, and Text Cleaning Subsystem employs Natural Language Processing (NLP) and Machine Learning techniques to extract meaningful data from news articles at Texas A&M University. The Text Processing, NLTK, and TextBlob components are a methodical approach to article analysis, involving processes such as standardizing text, substituting characters, eliminating non-alphabetic elements, removing stopwords, and lemmatization.

2.6. Future Planning

Looking ahead, future developments aim to capitalize on the model's efficacy in handling shorter texts. There is a possibility of splitting each article’s content into individual sentences, obtaining a score for that sentence, and averaging each sentence score for each article. There is also code underway that aims to train a model based on the TextBlob Subjectivity Scores that have promising validation results (~70% accuracy). From this, the project could be expanded to also be trained on the NLTK Sentiment scores. All of these efforts aim at improving validation and test accuracy for the model, since the more accurate the results, the better the data can be for the narrative curve analysis subsystem and the more sound the entire project is. The model has been run on numerous smaller datasets scraped from the scraper subsystem, but there is no good way to validate the data due to the dataset being so small. This aims to be fixed in the future when the scraper subsystem is capable of scraping hundreds, if not thousands of articles at a time.

NEWS Analytics

James Nieberding

SubSystem Report: Topic Model, and Narrative Curve Model

Change Record

Rev.	Date	Originator	Approvals	Description
1	12/03/23	James Nieberding		First Draft
2	4/29/2024	James Nieberding		Revised For End of Project Report

SUBSYSTEM REPORT:

TOPIC MODEL AND NARRATIVE CURVE MODEL

FOR

News Analytics

TEAM <12>

APPROVED BY:

_____

Project Leader Date

_____

John Lusher II, P.E. Date

_____

T/A Date

Table of Contents

List of Tables

List of Figures

2.1 Subsystem Introduction

The Topic Model is designed to group the articles based on the single words as well as grouping of words when analyzing the whole article. Using Gensim's LDA model library, we can find these different topic groupings as well as assign the probability that articles are related to a specific topic. This is used to separate articles based on their contents so we can send these different topics to the narrative curve model to be shaped.

The Narrative Curve model is designed using XGBoost's Regression model library to learn trends and how certain features relate to each other and how they affect sentiment over time. We can look into what the model is actually learning as well as the regression metrics to understand which features are leading to the peaks of sentiment at specific point in time.

2.2 Software Requirements and Documentation

The code developed for these subsystems was developed to be used on personal computers, the ECEN Olympus servers, or Google Colab' servers. It is written in Python and utilizes many different python libraries to get its output.
It uses XGBoost, NLTK, Pandas, Gensim, Sklearn, and Numpy to name a few.

Table 1: Reference Documentation

Document Number	Revision/Release Date	Document Title
1	3.11.5	Python Library Reference
2	41ce8f28	XGBoost Documentation
3	3.8.1	NLTK Documentation
4	2.1.3	Pandas Documentation
5	4.3.0	Gensim Documentation
6	0.21.3	ScikitLearn Documentation
7	1.26	Numpy Documentation

2.3 Subsystem Details

2.2.1. Topic Model

This part used Gensim's LDA model library to sort documents into topic buckets which can be used by the Narrative Curve model. It first takes the cleaned articles coming from the text cleaner and analyzer and checks for any extra stop words that it might have missed. It then creates a bag of words which essentially gives the words specific ID numbers which can be used by the model.

LDA models are probabilistic models which look at these IDs and find where the documents converge to create the specific topics. The common evaluation metric of this model is its coherence score which essentially tells the user how easily interpretable the topics are to a user. The model can be tuned using hyper parameters so that it can maintain the highest coherence score associated with the dataset. Once these topics are made, we check the probability that a certain article is associated with a topic and tune the threshold minimum probability so that the topic bucket it creates is actually over that specific topic. Once the topic buckets are created then it can be sent to the narrative curve analyzer.

2.2.2. Narrative Curve Model

This part of the system uses XGBoost's Regression model library to understand and learn the relationship between the feature scores that are coming out of the text analyzer. This model was trained using a dataset with 2225 articles that were cleaned and analyzed by the cleaner and analyzer. It was then sorted into the topics that the topic model found in the dataset. The articles in this dataset however did not have any dates associated with them. The articles were sorted in such a way that there was a general trend that the sentiment followed so the model could learn the trend and make predictions as a proof of concept.

2.4 Subsystem Validation

2.4.1. Topic Model

In order to validate the topic model and the methods to get an accurate model, a dataset with 2225 articles with 5 predefined topics was given to the model. The hyper parameters like eta, alpha, and the number of topics were tuned to maximize the coherence score of the LDA model.

The number of topics was tuned first to find the best number of topics that fit the data set. Essentially, I ran a grid search on the different number of topics to find the best number. The results are shown below.

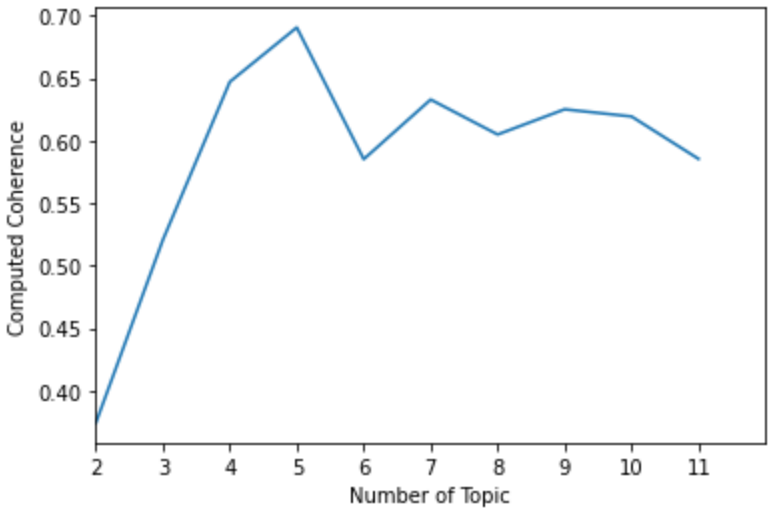


Figure 1. Coherence Measured over the different Number of Topics

Then I was able to tune other parameters like eta and alpha. Alpha is the document to topic density and essentially smooths over the weights of the topics so they fit the documents in the dataset better. Eta is the topic to word density and just determines the weights of the words so that they fit into the topics. This again used a grid search and the parameters associated with the highest coherence score were then found as seen below.

	A	B	C	D	E	F	G
1	alpha	eta	coherence	Highest Coherence	Max Number Index	Max Alpha	Max Eta
2	0.5	0.5	0.69492498	0.699969331	83	0.62	0.92
3	0.5	0.53	0.69492498				
4	0.5	0.56	0.69492498				
5	0.5	0.59	0.69492498				
6	0.5	0.62	0.69492498				
7	0.5	0.65	0.69492498				
8	0.5	0.68	0.69492498				
9	0.5	0.71	0.69492498				
10	0.5	0.74	0.69492498				
11	0.5	0.77	0.69492498				
12	0.5	0.8	0.69492498				
13	0.5	0.83	0.69492498				
14	0.5	0.86	0.695104507				
15	0.5	0.89	0.695104507				
16	0.5	0.92	0.695104507				
17	0.5	0.95	0.695104507				
18	0.5	0.98	0.695104507				
19	0.53	0.5	0.699789804				
20	0.53	0.53	0.69492498				
21	0.53	0.56	0.69492498				
22	0.53	0.59	0.69492498				
23	0.53	0.62	0.69492498				
24	0.53	0.65	0.69492498				
25	0.53	0.68	0.69492498				
26	0.53	0.71	0.69492498				
27	0.53	0.74	0.69492498				
28	0.53	0.77	0.69492498				
29	0.53	0.8	0.69492498				
30	0.53	0.83	0.69492498				
31	0.53	0.86	0.69492498				
32	0.53	0.89	0.695104507				
33	0.53	0.92	0.695104507				
34	0.53	0.95	0.695104507				

Figure 2. Coherence Grid Search over Alpha and Eta Hyper Parameters

Once all these parameters are found we can then look into the documents and see what topics they fit into. The data frame created when the sorted documents are created also contained the predefined topic that came with the dataset.

Below is a screenshot of the topics that were found in the dataset.

```
Topics Found
Topic 0:
 (0, '0.013*"film" + 0.008*"best" + 0.007*"award" + 0.005*"music" + 0.004*"star" + 0.003*"year" + 0.003*"actor" + 0.003*"song"
+ 0.003*"band" + 0.003*"album"')

Topic 1:
 (1, '0.007*"company" + 0.006*"firm" + 0.004*"mr" + 0.004*"share" + 0.003*"bank" + 0.003*"deal" + 0.003*"court" + 0.003*"china"
+ 0.003*"site" + 0.003*"sale"')

Topic 2:
 (2, '0.017*"mr" + 0.007*"government" + 0.006*"party" + 0.006*"labour" + 0.006*"election" + 0.005*"blair" + 0.005*"people" + 0.
005*"minister" + 0.005*"tory" + 0.004*"brown"')

Topic 3:
 (3, '0.008*"people" + 0.006*"mobile" + 0.006*"game" + 0.005*"phone" + 0.005*"technology" + 0.004*"service" + 0.004*"user" + 0.
004*"year" + 0.004*"computer" + 0.003*"net"')

Topic 4:
 (4, '0.006*"game" + 0.005*"england" + 0.005*"player" + 0.004*"win" + 0.004*"match" + 0.003*"club" + 0.003*"play" + 0.003*"fina
l" + 0.003*"world" + 0.003*"ireland"')
```

Figure 3. Topics Found in the Validation Dataset

We see that topic 0 is associated with words about entertainment, topic 1 is related to business, topic 2 is related to politics, topic 3 is related to tech, and topic 4 is related to sports. We can then look into the sorted topic CSVs created to see if the predefined topics match the topics the model found.

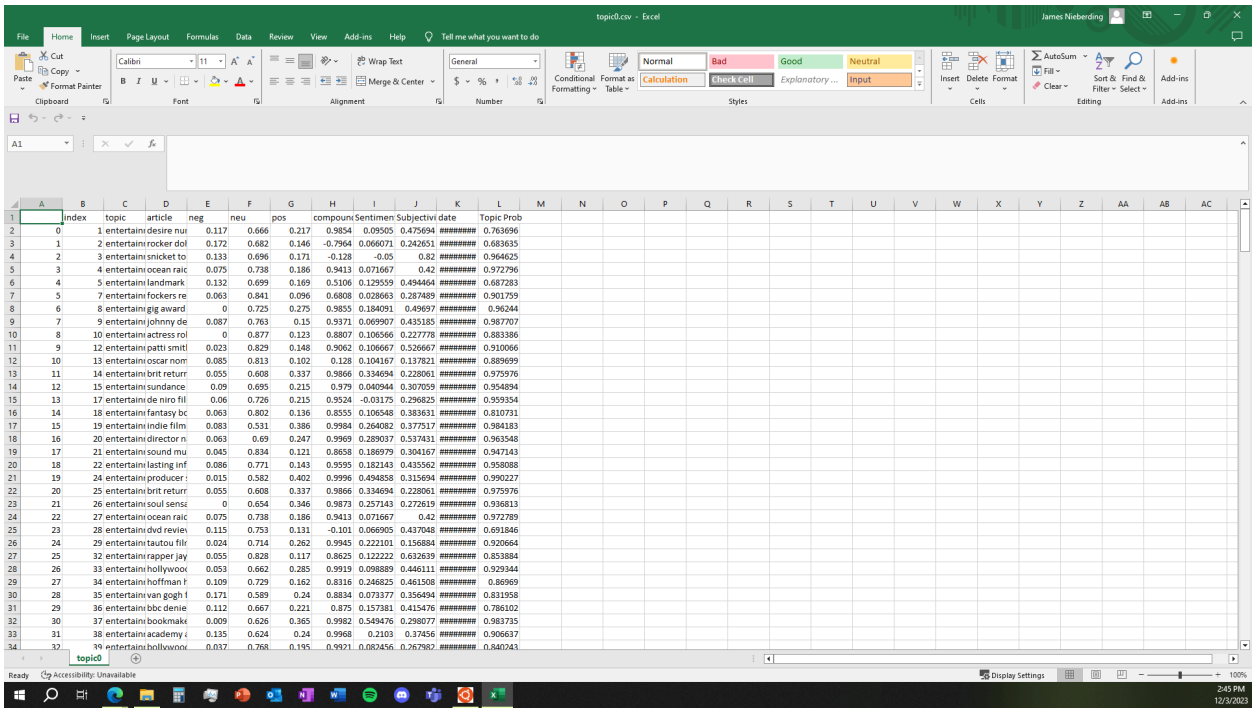


Figure 4. Topic 0 CSV results

The topic 0 CSV does in fact contain articles that relate to entertainment.

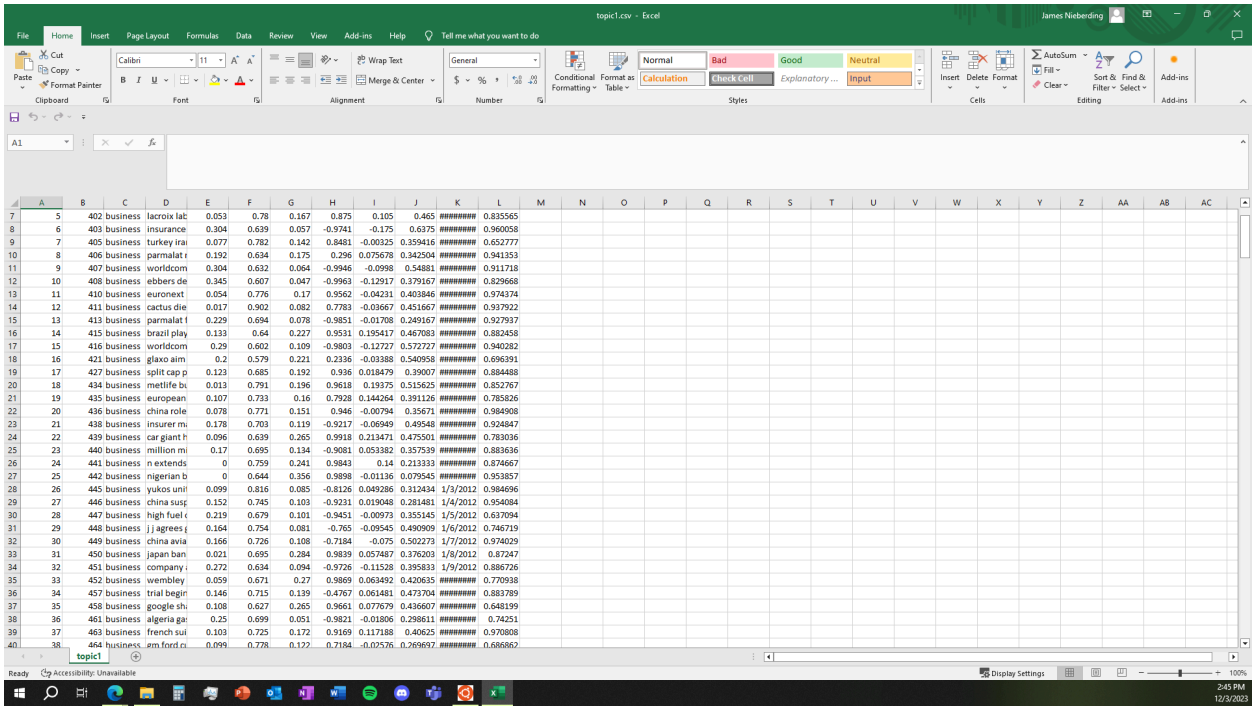


Figure 5. Topic 1 CSV results

The topic 1 CSV does in fact contain articles that relate to business.

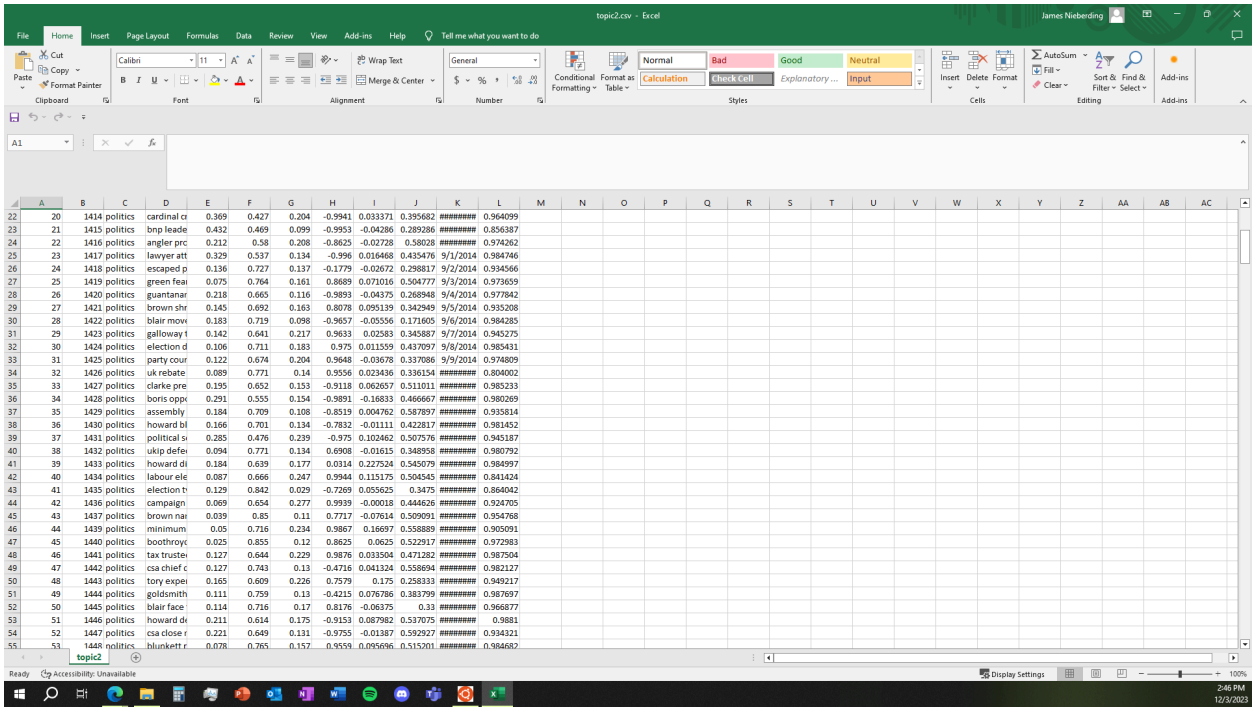


Figure 6. Topic 2 CSV results

The topic 2 CSV does in fact contain articles that relate to politics.

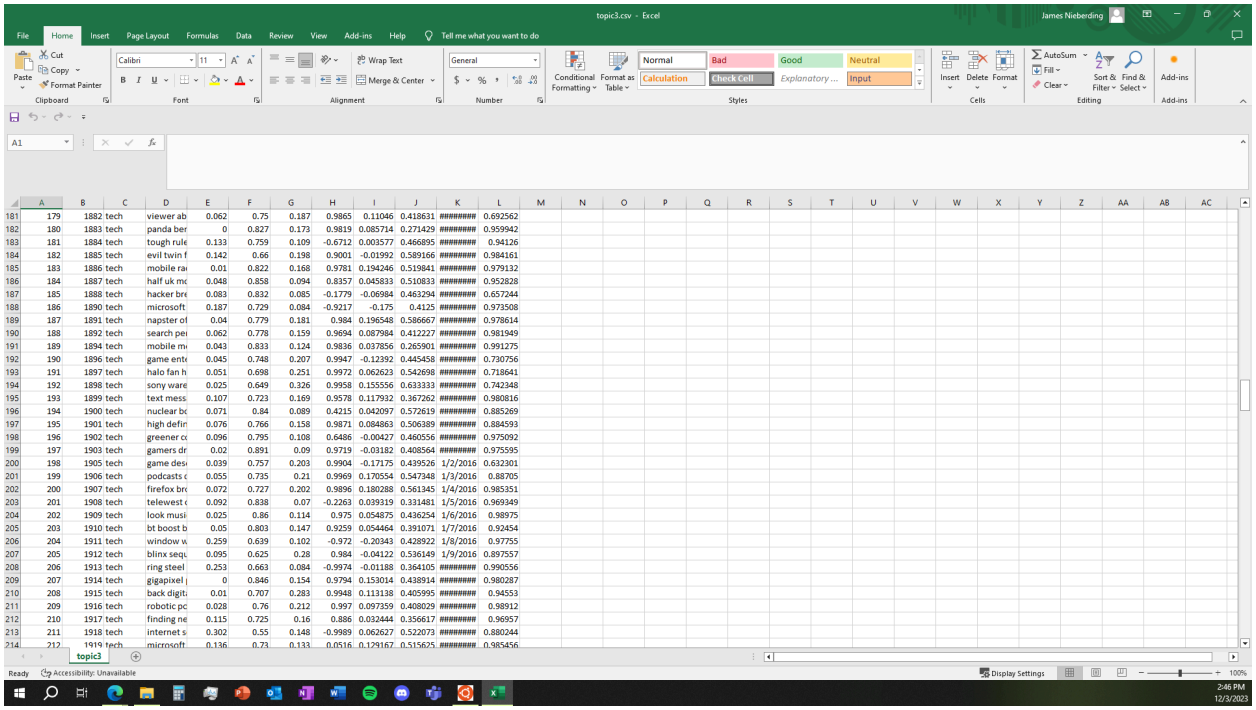


Figure 7. Topic 3 CSV results

The topic 3 CSV does in fact contain articles that relate to tech.

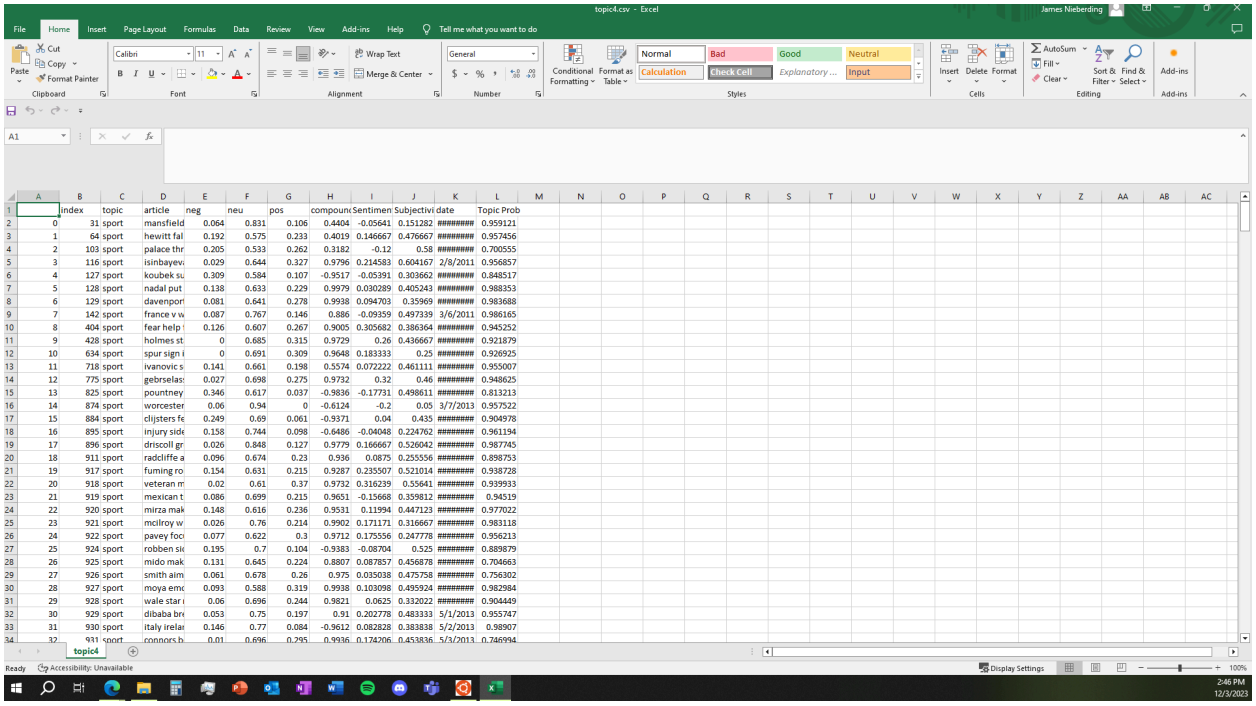


Figure 8. Topic 4 CSV results

The topic 4 CSV does in fact contain articles that relate to sports.

Another thing to note is the topic model is able to find topics that at face value you wouldn't have associated with a topic. An example, in the sorted topic 0 csv which is about entertainment, there is an article that was predefined to be in the tech topic however it

talked about the BCC and the national theater which is related to entertainment. So, the topic model is able to find the underlying themes in the articles and group them accordingly.

2.4.2. Narrative Curve Model

As stated before, the dataset with 2225 articles had to be sorted so that there was a general trend that the model would be able to learn and find. Below is a sort for topic 0 that I used to define my methods to validate my model. The y-axis represents the sentiment of the topic 0 and the x-axis represents time over about 550 days. The first 85% was used for training and testing, and the last 15% was used for validation. Of the 85%, 15% of that was used for testing and 85% was used for training.

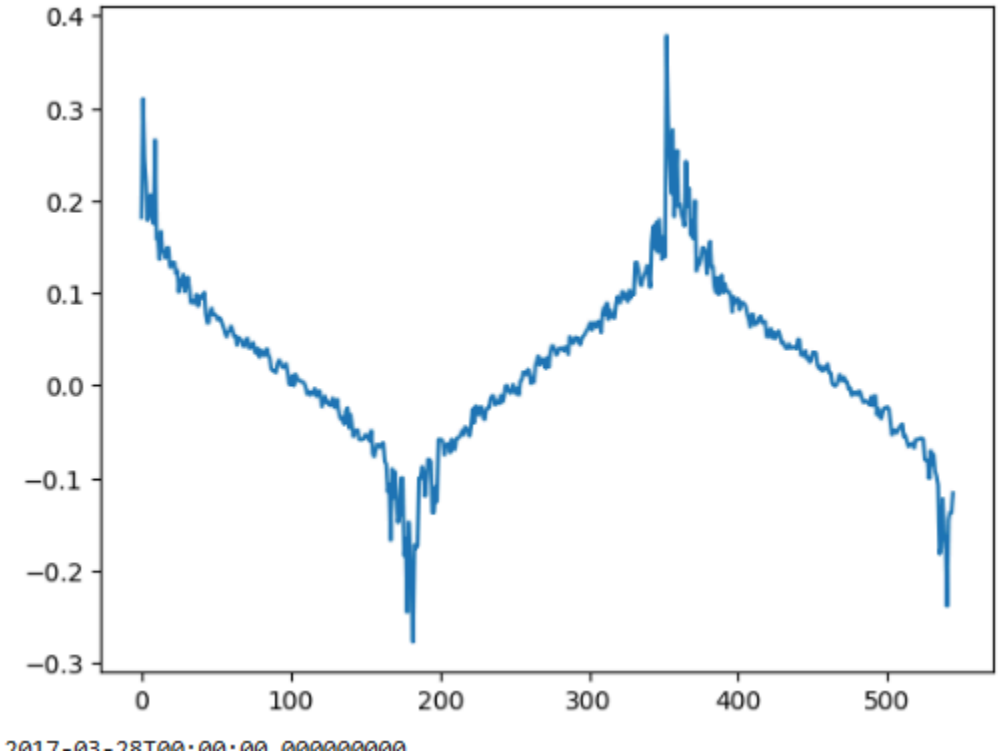


Figure 9. Sentiment over Time Graph Example

First, a board grid search was run to find rough ballpark values for the best hyper parameters. This is just so a model that is more accurate and actually has reasonable results. Then the relationship between the features was evaluated to see what combination of features resulted in the best error and R^2 scores. Once this was done, then the hyper parameters could be more finely tuned and the genuinely best parameters can be found.

	Broad Grid Search						Feature Tuning					
Run	Results	R2 Score	R2 Pred				Feature Vector	R2 Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	
1	learning=0.1833 depth=5 child=1 estimators=200	0.76178357		This one did not have topic prob or compound score			No Features Dropped	0.8362				
2	learning=0.19 depth=6 child=1 estimators=400	0.74480803		This one also did not have topic prob and compound			Topic Prob	0.83615604				This looks exactly like the No Feature drop
3	learning=0.27325 depth=6 child=1 estimators=600	0.711423631	0.79402	This was using all features one thing to note is child and estimators were on edge		Features Tuning will use the best params found using grid search to find relationship between features	Date w/ Polarity Scores	0.798493876	0.02820237	0.001683737	0.041033366	
4	learning=0.27325 depth=6 child=1.01 estimators=800	0.691430065	0.83616	estimators was bottomed out			Only Date	0.882127675	0.021950905	0.000984913	0.031383324	It's literally better in every metric from other test runs so far (r2, mae, mse)
5	learning=0.27325 depth=6 child=1.01 estimators=850	0.691430066	0.83616				Date w/ Subjectivity and Topic Prob	0.763587762	0.030443563	0.001975404	0.044445521	

Figure 10. Broad Grid Search and Feature Tuning Results

```
TESTING SCORES
Mean Absolute Error: 0.016750196939633397
Mean Squared Error: 0.000737764806834891
Root Mean Squared Error: 0.027161826279447614
R2 Score: 0.9191552430391913
Adjusted R2 Score: 0.9101724922657681


VALIDATION SCORES
Mean Absolute Error: 0.015895197147519365
Mean Squared Error: 0.0006557182528436617
Root Mean Squared Error: 0.025606996169868532
R2 Score: 0.7457062246184568
Adjusted R2 Score: 0.7174513606871742
```

Figure 11. Regression Metric Scores

Above the error scores of the model when predicting the sentiment using the testing and validation feature values. We see that the average error scores associated with both the testing and validation set are very low. Looking at the R2 scores, we can look into how accurate the model actually is. This metric tells us how well the model would be able to predict with unseen data. We see that the validation R2 score is very good for the testing set but dips for the validation set. This is as expected because prediction upon unseen data will not be as accurate.

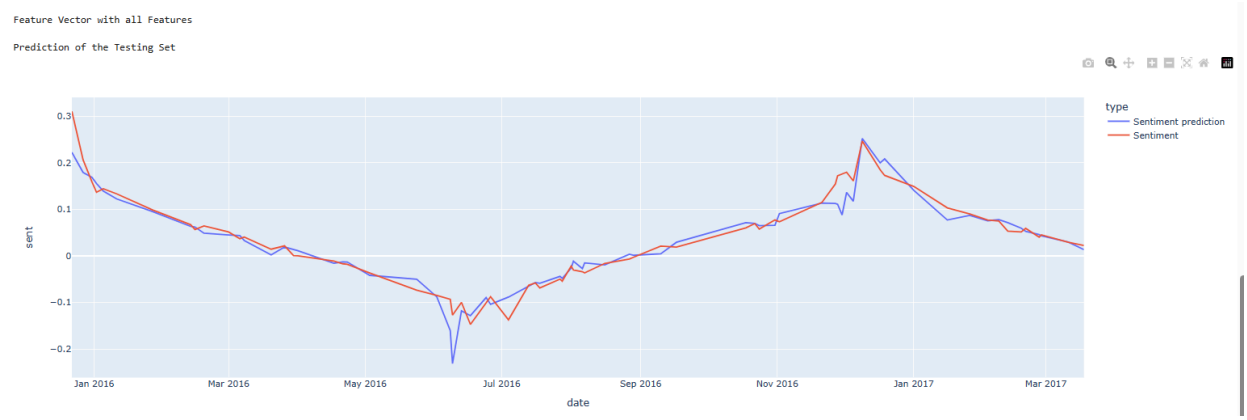


Figure 12. Testing Set Sentiment Prediction Over Time

Above is the graph of the predicted sentiment on the testing set. We see that around December there is relatively high sentiment and the model was able to predict it.

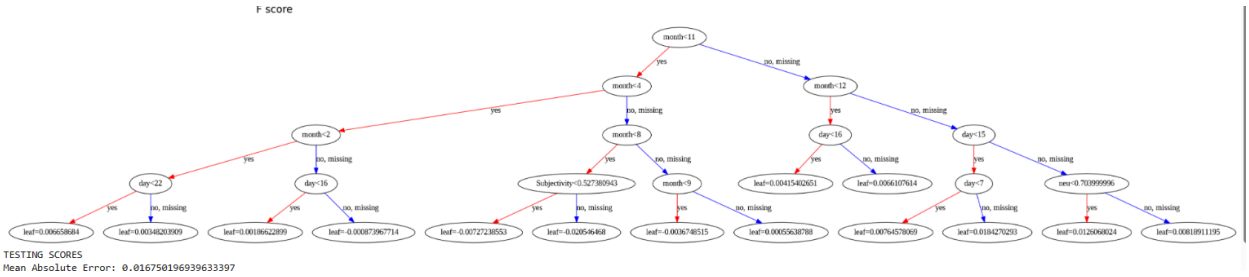


Figure 13. XGBoost Tree Graph of Model

Above is the actual tree branch model that is created. We can look into the branches to find out how the model is creating its predictions.

Figure 14. December Branch

Here is the branch that is associated with December and what we find is that the leaf weights for this branch are some of the highest weights in the model. This shows us that the model was able to learn that there was high sentiment during this period of time and gave it a higher weight respectively.

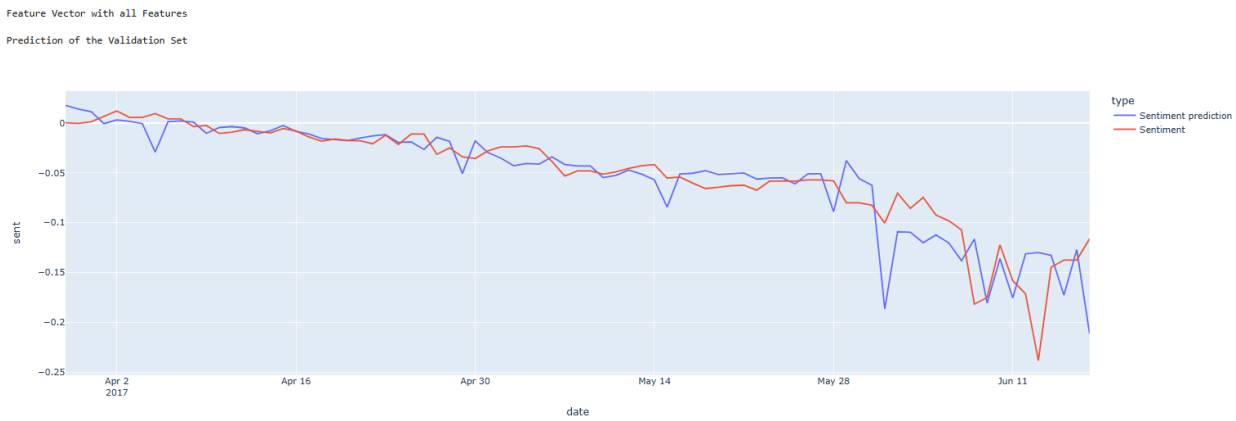


Figure 15. Validation Set Sentiment Prediction Over Time

Above is the prediction results on the validation set over time. We see that it starts off fairly well with its predictions but towards the end it loses some accuracy when dealing with more extreme cases.

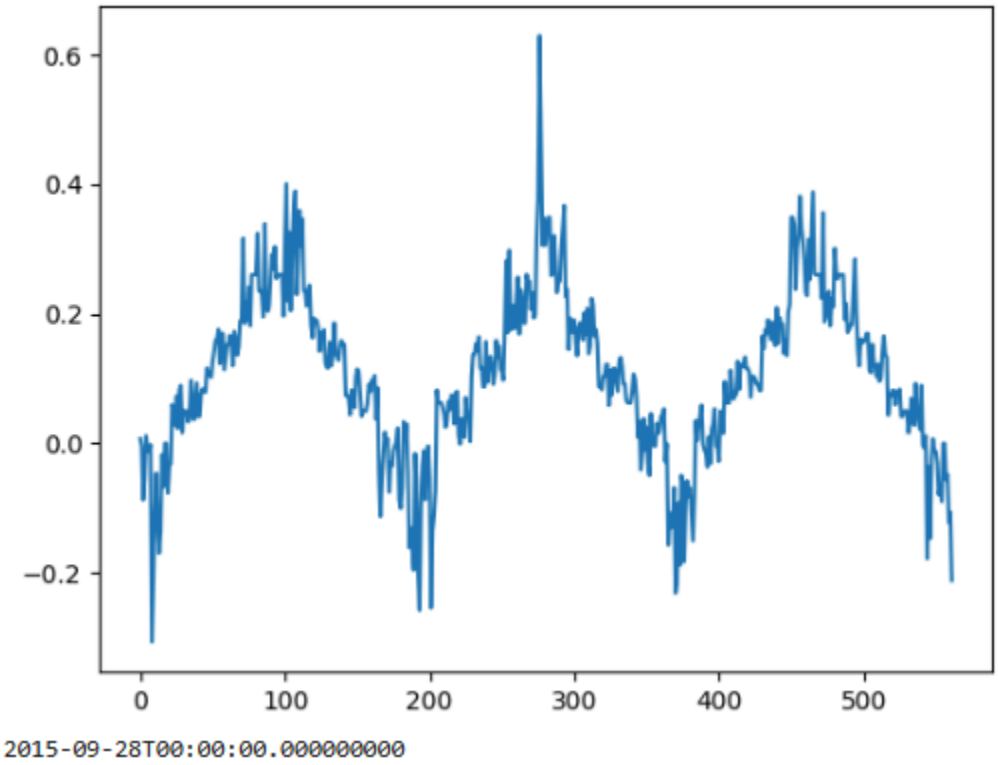


Figure 16. Nosy Sentiment Curve Over Time

Above is another narrative curve that was used to evaluate the effectiveness of the model. This curve was more noisy in order to prove that this method could handle noisier data.

```
TESTING SCORES
Mean Absolute Error: 0.04500652220879023
Mean Squared Error: 0.003657069579244204
Root Mean Squared Error: 0.06047370981876508
R2 Score: 0.7492874865033448
Adjusted R2 Score: 0.7225448183970349


VALIDATION SCORES
Mean Absolute Error: 0.03582342587795931
Mean Squared Error: 0.0021987254432942263
Root Mean Squared Error: 0.04689056880966818
R2 Score: 0.8170580930440138
Adjusted R2 Score: 0.7975442896353753
```

Figure 17. Regression Metric Scores on Noisy Data

This is the mean error scores as well as the R2 scores associated with this curve.

Figure 18. Sentiment Prediction on Testing and Validation Set Over Time

Here is the sentiment prediction of the model over time on the testing and validation set. The model can follow the trends of the curve and has decent scores when it comes to its prediction. The model can only be improved and when it comes to integration, we will be gathering more data points over a longer period of time so that the model can gather as much information as it can to handle more extreme values.

2.5. Conclusion

To conclude, the results of this report have been a proof of concept and the methods developed this semester will be used next semester to create the whole model over real news data. The methods for the topic model have been developed and successfully worked on a dataset that already had 5 predefined topic buckets. It was able to find these topics and create its own topic buckets. The narrative curve model was able to also find the general trends of sentiment over time using the methods developed and produced a fairly accurate model.

2.6. Future Planning

Moving forward, this subsystem as well as the methods created will be used on the actual news data that scraper and analyzer will be producing. I found that the narrative curve model had a harder time handling extreme cases however with the scraper gathering as many news sources as possible the model should be able to have enough data points that it can handle these extreme cases.

Execution Plan

EXECUTION PLAN	10/2/23	10/9/23	10/16/23	10/23/23	10/30/23	11/6/23	11/13/23	11/20/23	11/27/23	12/3/23	Date
Status Update 1 (Project Introduction Presentation)											
Build Model Scrapper											
Find/Create Gensim Test Data										Not Started	
Build Gensim Framework										In Progress	
Acquire Valid Article Data for NLTK Parsing										Complete	
Expand Scrappers to target websites										Behind Schedule	
Complete NLTK text processing w/ Valid Data for Textblob											
Design Blitz											
Test Gensim Framework											
Begin testing Textblob Sentiment scores w/ Valid Data											
Acquire Preliminary data for Analyzer and Modeler											
Create XGBoost Model											
Find/Create Training and Validation Data for XGBoost Model											
Database Creation											
Status Update 2 (Project Update Presentation)											
Format XGBoost Output to Narrative Curves											

Verify Textblob outputs Subjectivity/Sentiment scores w/o errors											
Database Verification and Migration to Olympus											
Connect from Scraper to Olympus Database											
Test and Validate XGBoost Model											
Verify Scraper Connection Populate Database											
Create tensorflow model for Emotion Recognition on articles											
Apply model to extract numerical values for emotions on articles											
Status Update 3 (Final Presentation)											
Subsystem Demonstration											
Final Report											

Validation Plan

Paragraph #	Test Name	Success Criteria	Methodology	Status	Responsible Engineer(s)
3.1.1	Scrapper Deployment	Data acquired match target website	Manually verify that the data comes from the target	TESTED	Alexander Corpstein
3.1.1	Scrapper Data Verification	Data acquired matches the HTML of the target website	Manually verify that the data extracted matches the target HTML	TESTED	Alexander Corpstein
3.2.3.2	Input Data Format	Data is formatted for insertion	Verify that the data is correctly formatted for insertion into the database	TESTED	Alexander Corpstein
3.2.3	Scrapper to Database Connection	Scraper successfully sends data to the Database	Load data from the Scrapper into the database then request the data from the database and verify the data	TESTED	Alexander Corpstein
3.1.1	Database Creation	Database is successfully setup	Verify that the .db file is create in the correct directory in the Olympus server	TESTED	Alexander Corpstein
3.1.1	Database Storage	Database successfully stores and outputs data	Load preselected data then request it and verify completeness	TESTED	Alexander Corpstein
3.2.3.1	Olympus Connection	All project members can access Olympus server	Ensure that all team members can access the team directory	TESTED	Alexander Corpstein
3.1.2	NLTK Parsing/ Cleaning	Successfully parse/clean data from database/internet	Remove stop words, punctuation and formatting to prepare for sentiment/subjectivity analysis	TESTED	Alexander Corpstein
3.2.3.2	Communicate with SQL database	Successfully able to pull SQL data out of the database for analyzing	Create a program that is connected to the server and pulls test data to Pandas	TESTED	James Nieberding

3.1.3	SQL DB to Pandas	SQL DB data is converted into Pandas Dataframe that is usable for Gensim	Manually feed data that is in the database format to Pandas to test the output	TESTED	James Nieberding
3.2.1.4	Initial Textblob testing	Successfully analyze multiple sentences of text data for sentiment and subjectivity without errors	Use any large text article with predetermined scores for the sole purpose of testing scoring output	TESTED	James Nieberding
3.2.1.4	Textblob testing using actual data	Successfully analyze text data from NLTK processing without any errors	Using NLTK processed data from earlier testing, produce Sentiment/Subjectivity scores for real articles	TESTED	James Nieberding
3.1.3	Gensim Initial Framework test	Gensim successfully groups predetermined articles into topics	Create test data with predetermined data points that fit into certain topics and feed it into Gensim algorithm	TESTED	James Nieberding
3.1.3	Gensim Final Framework test	Check topics and articles associated with each topic to ensure it sorts them correctly	Feed the algorithm unknown data and perform topic modeling	TESTED	James Nieberding
3.2.1.3	Trained XGBoost model	Model is able to correctly accept the validation set and create predictions	Create a validation data set and feed it into the model	TESTED	James Nieberding
3.2.1.3	XGBoost Accuracy	The model is able to replicate the shape of the sentiment narrative curves up until present day	Feed the model data up until 6 months ago	TESTED	James Nieberding

NEWS Analytics

Alexander Corpstein
James Nieberding

INTEGRATED SYSTEM REPORT

Change Record

Rev.	Date	Originator	Approvals	Description
1	4/29/23	James Nieberding		Revised For End of Project Report

INTEGRATED SYSTEM REPORT

FOR

News Analytics

TEAM <12>

APPROVED BY:

_____

Project Leader Date

_____

John Lusher II, P.E. Date

_____

T/A Date

Table of Contents

List of Tables

List of Figures

1.1 Overview

Integrating the system involved connecting both the news scrapers/text cleaners and the topic model/narrative curve model to a database that was created on the TAMU ECEN Olympus servers. Integration mainly involved inputting the data from scraper in a way that the topic model could then analyze the dataset and pass the data off to the narrative curve model.

2.1 Execution

Before integration began, a team member left and had developed the text cleaner and sentiment/subjectivity analyzer code. The code developed was split and the text cleaner was given to the team member with the web scraper while the sentiment/subjectivity analyzer was given to the team member with the topic model. The code was added to the pipeline of the other subsystems and validated to ensure it still worked properly.

Once this was done, then integration could begin. Below is the execution plan outlining the different checkpoints and tasks done over the integration process.

Table 1: Execution Plan

EXECUTION PLAN	1/22/24	1/29/24	2/5/24	2/12/24	2/19/24	2/26/24	3/4/24	3/11/24	3/18/24	3/25/24	4/1/24	4/8/24	4/15/24	4/22/24	4/29/24
Rescope Project for New Semester															
Add Text Cleaner to Scraper															
Add Sentiment/Subjectivity to Narrative Model															
Validate Narrative curve model with sentiment analyzer															
Verify Scrapy works with NLTK cleaner															
Bi-Weekly Update 1															
Connect database to Model															
Fill Database with Entries															

Test connection from Database to Model															
Bi-Weekly Update 2															
Tune Topic Model Hyper Parameters															
Validate test data from database works with Model															
Bi-Weekly Update 3(Blitz)															
Train Model with data from database															
Database Cleanup															
Validate Model															
Bi-Weekly Update 4															
Database Finalization															
Bi-Weekly Update 5															
Creation of Technical Report for Sponsor															
Create graphical output based on analyzed data from ML model			In Progress												
Final Design Presentation			Complete												
Final Project Demonstration			Behind Schedule												
Final Report			Not Started												

3.1 Validation

3.1.1 Scraper to Text Cleaner To Database

With the Text Cleaner added to the pipeline of the scraping this semester it allowed for a streamline of the article data to the database. In Figure 1 we can see an example of how the text cleaner worked. The data scraper from a website was used by the NLTK to remove stopwords and lemmatize the data. As well all the data was made uniform by making them lowercase and removing all punctuation and nonletters. A special case was used for “Texas A&M” to make it tamu instead to capture that information for the models.

	Headline	Date	Article	Publisher	Website
	qatar	Filter	Filter	Filter	Filter
1	tamu student faculty qatar slam decision close middle east campus	02/12/2024	student facult...	texas tribune	https://...

Figure 1. Example Text Cleaning to Database

To validate the text cleaner articles were examined before and after going through the pipeline to ensure that all the criteria were met for use by the models.

3.1.2 Topic Model

To validate the topic model, I used the coherence score associated with the model to determine the best fit model.

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} \text{NPMI}(w_i, w_j)^{\gamma} \right\}_{j=1,\ldots,|W|} \tag{3}$$

$$\text{NPMI}(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i, w_j)+\epsilon}{P(w_i) \cdot P(w_j)}}{-\log(P(w_i, w_j)+\epsilon)} \right)^{\gamma} \tag{4}$$

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \tag{5}$$

(iv) The final coherence score is the arithmetic mean of all confirmation measures ϕ.

Figure 2. LDA Model Coherence Score Formula

Above are the formulas that are used to calculate the coherence score of a topic model. Simply put the score is the mean of a cosine similarity score of word vectors that are found within the dataset. These formulas are fairly robust which provides a score that is normalized and quantifies how a human would correlate and group topics.

3.1.3 Narrative Curve Model

To find the best fit narrative curve and model, a grid search that iterates through parameters like model layer depth, learning rate, minimum child weight, number of training iterations and lambda. With each iteration, the R^2 score was calculated and saved to be compared to other iterations. The R^2 score or coefficient of determination tells us whether a regression model can predict the outcome of a dependent variable.

There were certain news articles that were scraped from as early as 2009 and were not terribly important to the near past. Years that didn't have more than 50 articles were pruned from the model's input feature vector to better understand the correlation of sentiment in recent years. Grid searches were then performed and R^2 scores were saved and compared to the previous results until the best fit model was found.

4.1 Performance

4.1.1 Topic Model

The best fit topic model for the data set we scraped was a topic model with 7 topics with a coherence score of 66.1195%.

```
Coherence Score:  0.6611949390624747

Topics Found
Topic 0:
 (0, '0.012*"research" + 0.007*"energy" + 0.006*"science" + 0.006*"engineering" + 0.005*"researcher" + 0.005*"department" + 0.005*"project" + 0.005*"tec
hnology" + 0.004*"study" + 0.004*"nuclear"')
```

Research

```
Topic 1:
 (1, '0.021*"student" + 0.018*"tamu" + 0.014*"said" + 0.013*"university" + 0.009*"school" + 0.008*"year" + 0.007*"program" + 0.006*"college" + 0.005*"ca
mpus" + 0.005*"new"')
```

Campus Policy News

```
Topic 2:
 (2, '0.028*"said" + 0.006*"health" + 0.006*"like" + 0.005*"also" + 0.005*"one" + 0.005*"people" + 0.005*"time" + 0.005*"year" + 0.004*"could" + 0.004
*"get"')
```

General Campus News

```
Topic 3:
 (3, '0.062*"def" + 0.033*"hou" + 0.033*"txst" + 0.022*"maria" + 0.014*"cortijo" + 0.014*"dzemeshkevich" + 0.014*"parreno" + 0.014*"pedemonti" + 0.014
*"slisane" + 0.011*"gabriella"')
```

Tennis Tournament

```
Topic 4:
 (4, '0.032*"tamu" + 0.014*"game" + 0.014*"aggies" + 0.013*"team" + 0.011*"point" + 0.010*"match" + 0.010*"season" + 0.010*"first" + 0.009*"double" + 0.
007*"win"')
```

Sports

```
Topic 5:
 (5, '0.022*"texas" + 0.017*"state" + 0.010*"law" + 0.010*"said" + 0.009*"court" + 0.007*"border" + 0.006*"federal" + 0.006*"abbott" + 0.005*"would" +
0.005*"republican"')
```

Government/Politics

```
Topic 6:
 (6, '0.009*"app" + 0.007*"device" + 0.006*"phone" + 0.006*"tablet" + 0.006*"e" + 0.006*"smart" + 0.005*"access" + 0.005*"thornton" + 0.005*"barclay" +
0.005*"network"')
```

Visualization Department/Art Exhibit

Figure 3. LDA Model Best Fit Model Topics

The coherence score is high and almost as high as the topic model with the 5 predetermined topics which was used as a proof of concept.

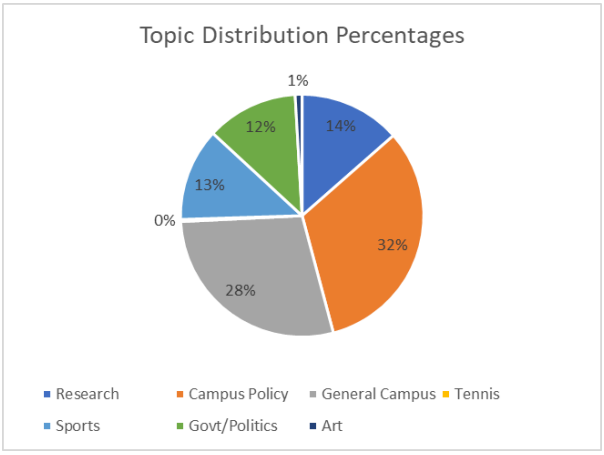


Figure 4. Topic Distributions

The distribution of articles was 32% being Campus Policy, 28% General Campus News, 14% Research, 13% Sports, and 12% Govt/Politics. There were 2 topics that had articles that were related to each other, the tennis tournament and art topic, but did not have enough to be used for the narrative curve model at the time.

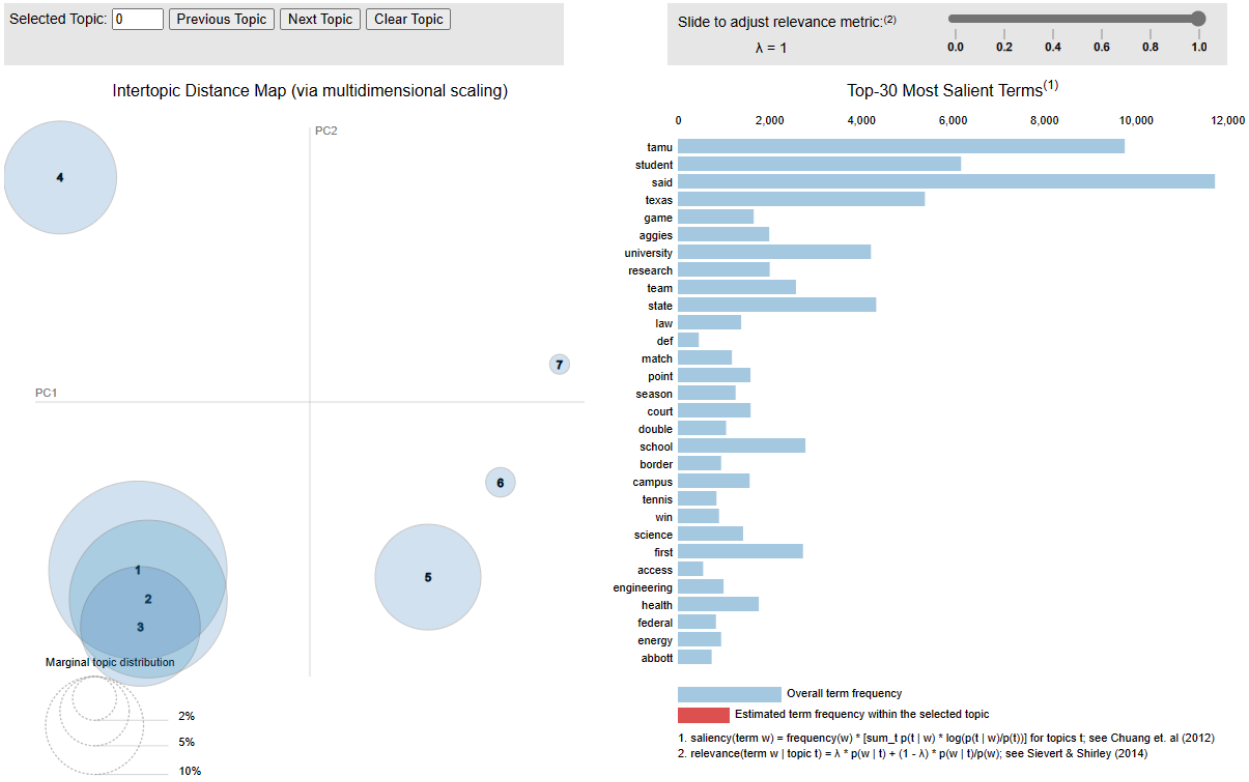


Figure 5. Intertopic Distance Map

Above is another representation of the different topics found in the dataset. This allows us to see how closely related certain topics are as well as the top words associated with the dataset and individual topics.

4.1.2 Narrative Curve

Multiple grid searches were performed on each topic by changing the feature vector or input training vector that the regression model saw. Parameters found and scores associated with the best parameters were recorded to keep track of the best regression model for a topic.

Table 2: Topic 0 Tuning Data

Topic 0		
	2017>	All
Run	topic0_tuning1	topic0_tuning2
Data Points	1079	1079
learning	0.1	0.01
depth	2	2
child	0.01	0.01
estimators	50	350
lambda	0.2	0
testing r2	0.146140501	0.204907179
validation r2	0.172146798	0.156056962

After this we could then look into the resulting prediction and scores associated with the best fit model.

Figure 6. Topic 0 Prediction of Testing Set

Figure 7. Topic 0 Prediction of Validation Set

This process was performed on all topics. Below are the results.

Table 3: Topic 1 Tuning Data

Topic1	
	All
Run	topic1_tuning1
Data Points	959
learning	0.1
depth	2
child	0.01
estimators	50
lambda	0
testing r2	0.470004965
validation r2	0.326897942

Figure 8. Topic 1 Prediction of Testing Set

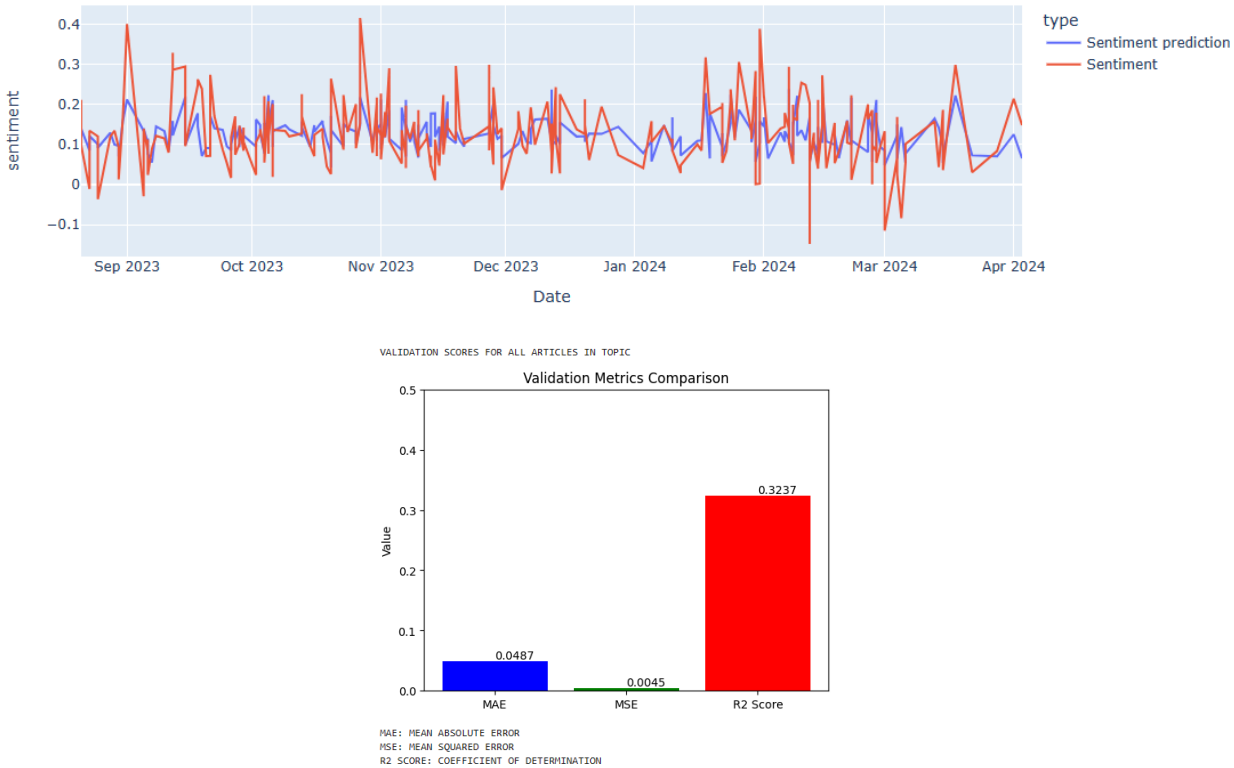


Figure 9. Topic 1 Prediction of Validation Set

Table 4: Topic 2 Tuning Data

Topic2		
	All	>2019
Run	topic2_tuning 2	topic2_tuning 3
Data Points	2458	4374
learning	0.01	0.01
depth	2	2
child	0.01	0.01
estimators	400	250
lambda	7.5	0.8
testing r2	0.30271308	0.340587552
validation r2	0.397301959	0.363916152

Figure 10. Topic 2 Prediction of Testing Set

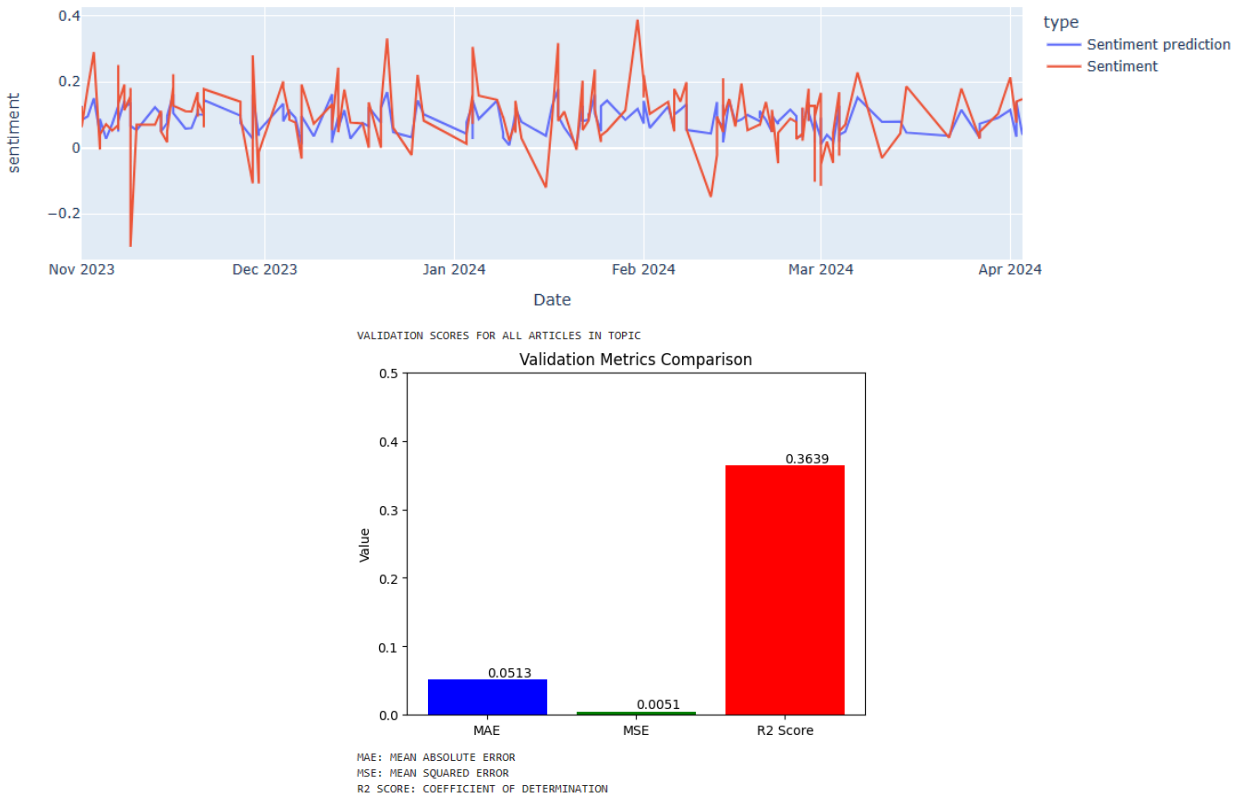


Figure 11. Topic 2 Prediction of Validation Set

Table 5: Topic 4 Tuning Data

Topic4			
	All	>2021	>2021
Run	topic4_tuning 1	topic4_tuning 2	topic4_tuning 3
Data Points	1079	1079	1511
learning	0.01	0.01	0.1
depth	3	4	7
child	0.01	0.01	0.01
estimators	250	200	200
lambda	0.4	0.5	0.4
testing r2	0.427	0.315014969	0.273643476
validation r2	0.353218022	0.198624106	0.253979445

Figure 12. Topic 4 Prediction of Testing Set

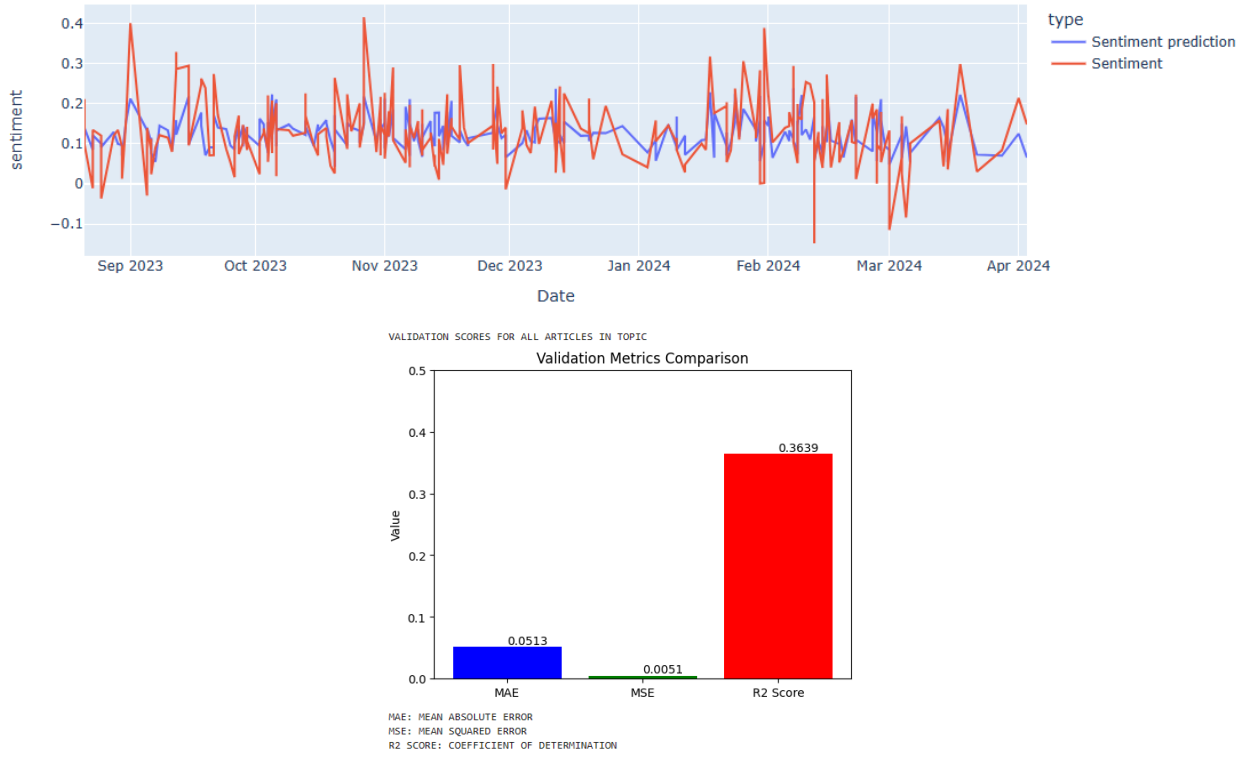


Figure 13. Topic 4 Prediction of Validation Set

Table 6: Topic 5 Tuning Data

Topic5					
	2022 and up (depth 7>)	2022 and up (depth 6<)	all (depth <6)	all (depth >7)	2018 and up (depth >7)
Run	topic5_tuning1	topic5_tuning2	topic5_tuning3	topic5_tuning4	topic5_tuning5
Data Points	5831	10799	10799	4535	4535
learning	0.01	0.01	0.01	0.01	0.01
depth	7	2	2	7	7
child	0.01	0.01	0.01	0.01	0.01
estimators	150	250	200	100	100
lambda	0.4	0	0.7	0.5	0.7
testing r2	0.14025098	0.047617938	0.249506888	0.135457342	0.019254745
validation r2	0.126412056	0.275859307	0.249030975	0.187973738	0.124385548

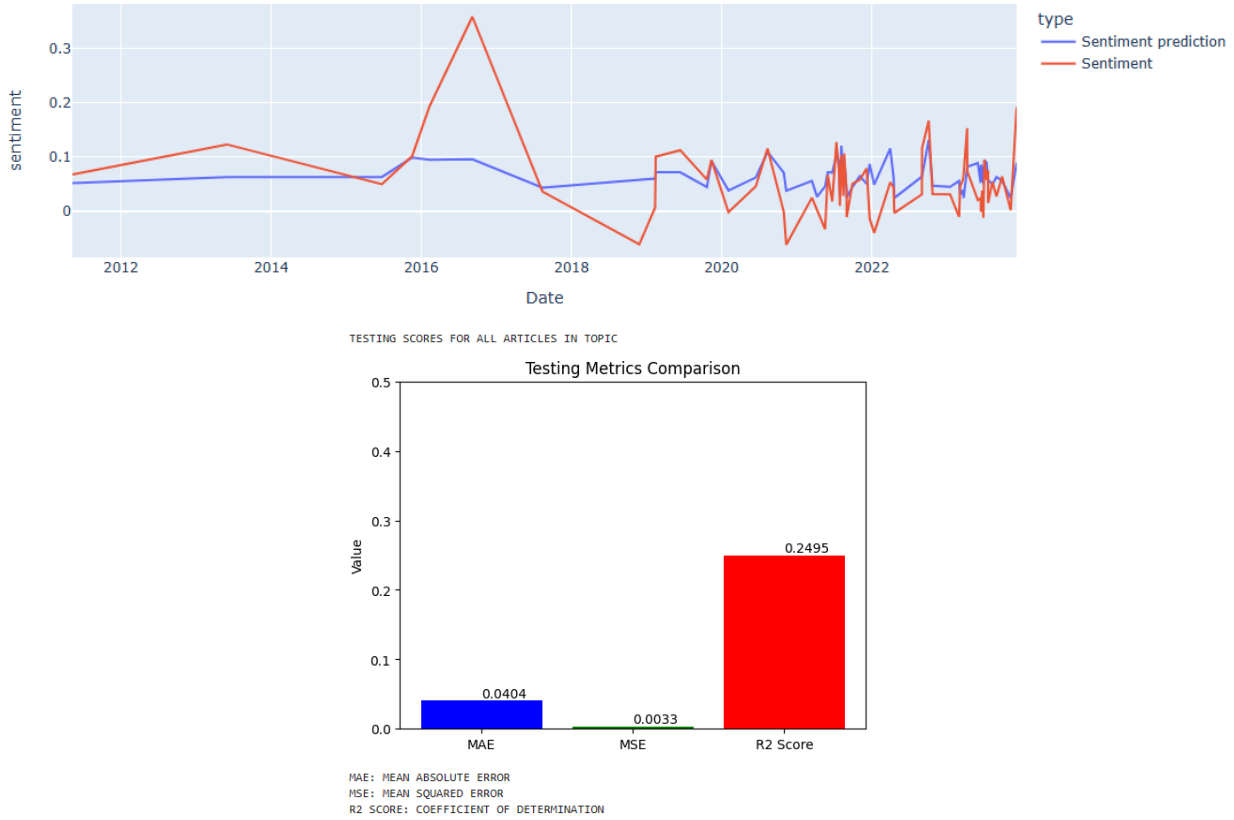


Figure 14. Topic 5 Prediction of Testing Set

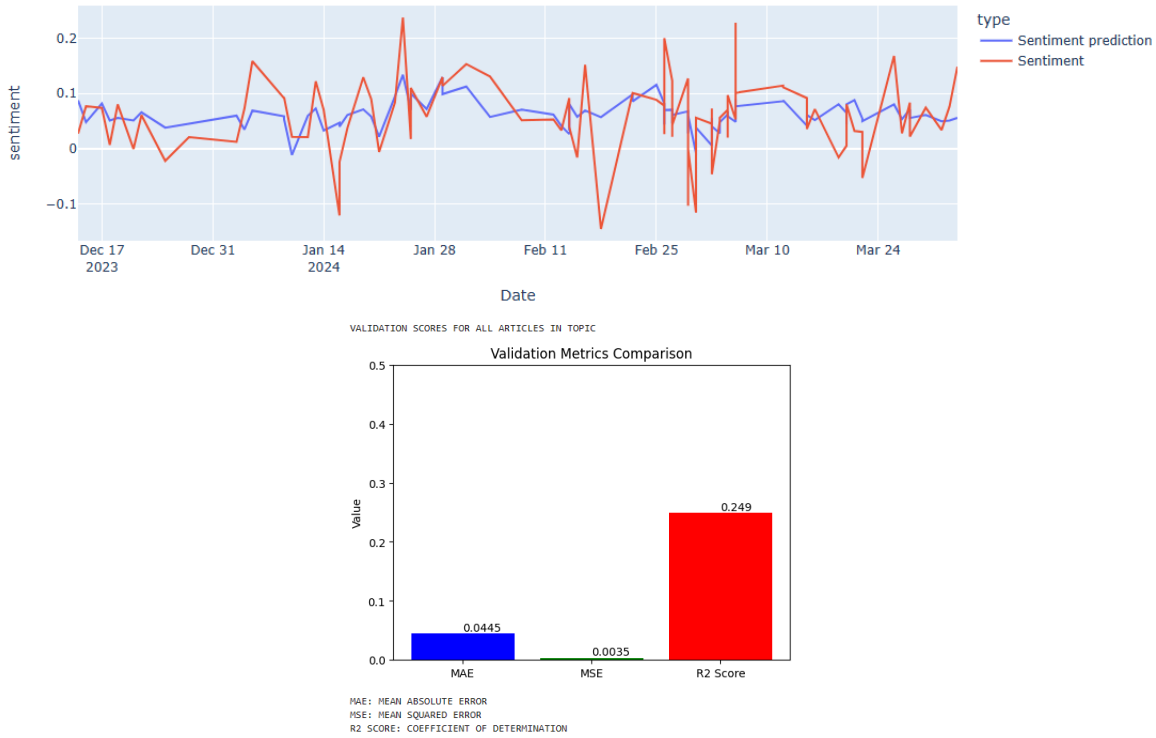


Figure 15. Topic 5 Prediction of Validation Set

5.1 Conclusions

5.1.1 Scraper Limitations

For this project only five scrapers were made leading to ~2500 articles scraped. The limitations for each newsite were different, but caused problems with the amount of articles that could be extracted. For example the Texas Tribune has a search function that only allows 10 pages of 10 articles meaning that only a max of 100 articles can be scraped from a single keyword search and not all selections would be articles, or relevant. The relevance of the articles scraped can be most seen in CNN where the search function did not filter general articles about Texas from ones about Texas A&M.

5.1.2 Database Limitations

The database being made with SQLite3 and hosted on the ECEN server Olympus caused problems with connectivity. To connect to the database the connection must go through a VPN to connect to TAMU then through an ssh connection to connect to Olympus then connect to the database. This caused a great deal of slowdown and time delay that caused all operations to the database to take longer than if another method such as Firebase was used.

5.1.3 Sentiment/Subjectivity Limitations

The sentiment and subjectivity analyzer used was the base model that is used by NLTK. As such any errors or problems that are inherent to the analyzer are not caught by the program.

5.1.4 Topic Model Limitations

The gensim's LDA topic model works best when the whole dataset is static as it needs all the documents to gather the best fit topic for a dataset. When new articles were added to the dataset, the topic model would drop in coherence score so new grid search would have to be performed and the model would have to be retrained. This made things difficult when the scraper continued to add more articles to the database.

The LDA model is also a probabilistic model to group articles together. When more articles would be added to the database and the model had to be retrained, there was no guarantee that the new topics would be similar to the previous topics. This withheld any progress that could be made on the narrative curve model as the topic distributions were changing with the growing dataset.

5.1.5 Narrative Curve Model Limitations

The models developed for each topic were able to achieve scores that were fairly accurate when analyzing raw data. It however is limited in the fact that most times

sentiment is highly variable when it comes to news articles. One author could write a piece a certain way that would be completely different from another author causing the sentiment analyzer to not have consistent trends. The models were also pretty shallow as a lot of the times the best depth parameter was just 2 which is not a very complex model. This goes along with the not being able to distinguish trends with other variables other than the few features that created the shallow models.