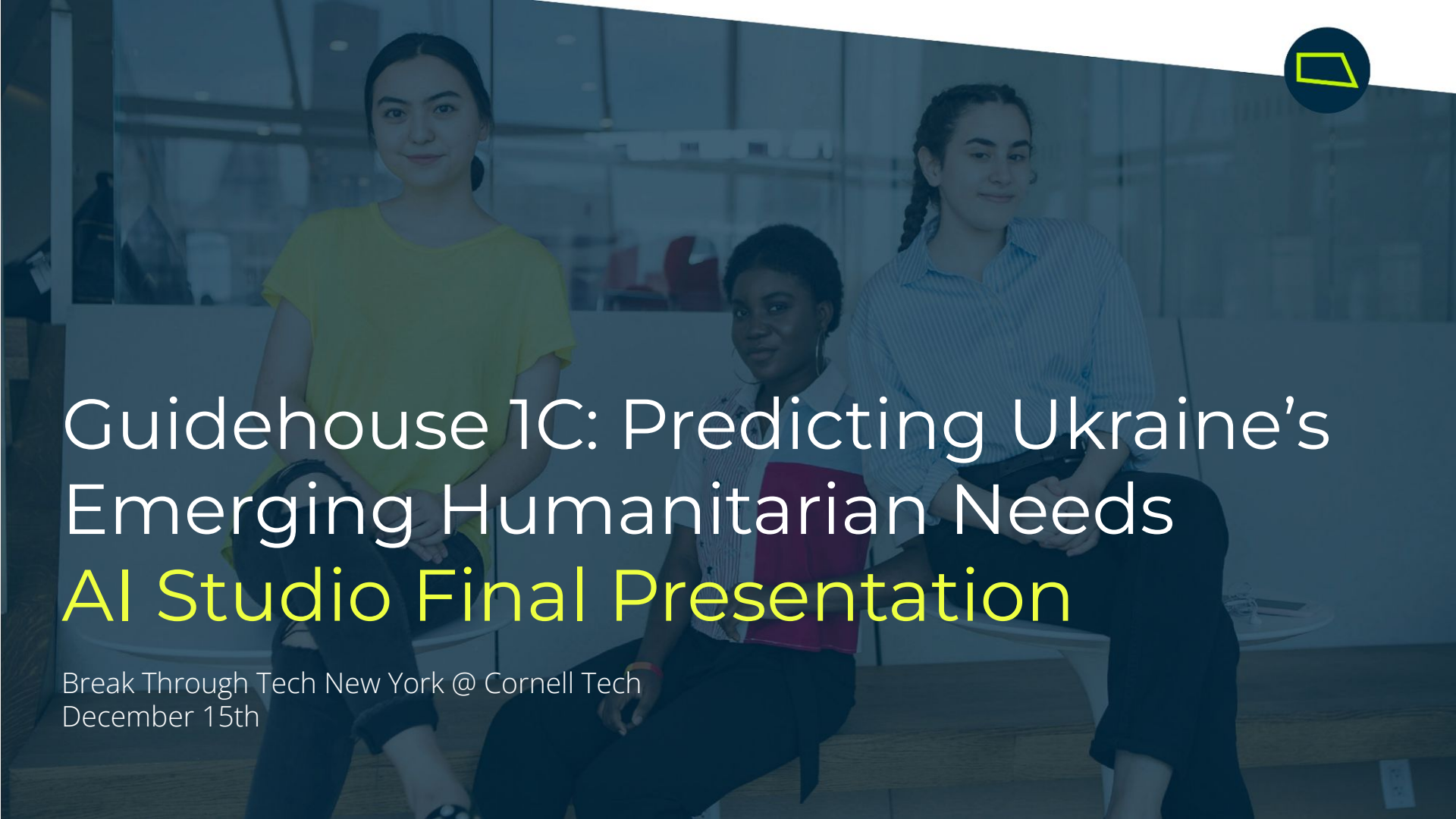


Guidehouse 1C: Predicting Ukraine's Emerging Humanitarian Needs

AI Studio Final Presentation

Break Through Tech New York @ Cornell Tech
December 15th

Introductions

Meet Our Team!

Niekelle Bloomfield
Rutgers University

Kerr Tan
New York University

Sowjanya Sritharasarma
Hunter College

Our AI Studio TA and Challenge Advisor

Sanjana Kaza
AI Studio TA

Darcy Watts
Challenge Advisor

Presentation Agenda

1. Project Overview & Business Understanding
2. Data Understanding & Exploration
3. Data Preparation & Preprocessing
4. Modeling & Evaluation
5. Key Insights & Final Thoughts

AI Studio Project Overview

Business Problem

- Our project tries to investigate the Ukraine humanitarian needs in different regions by adopting approaches including supervised and unsupervised learning.
- Many of Guidehouse's Defense and Security clients support humanitarian needs and people in need through the State Department, Department of Homeland Security, and Department of Defense.
- It is possible that solutions from this challenge could be adapted to address an actual client need in the future, such as the US Government.

Our Goals

1. Create **visualizations** to show the different **clusters** of people in need on the basis of oblasts (geographical regions within Ukraine), type of aid needed, and demographics of the people in need.
2. Build a **forecasting** model that displays trends in the aid needed over time based on how many people were injured or killed from the ongoing Russia-Ukraine conflict.
3. Build an NLP model that conducts **sentiment analysis** on tweets regarding the Russia-Ukraine conflict and analyzes whether or not the tweets are positive or negative.

Our Approach

Business Understanding

What is the business problem? What models can be implemented for this problem?

August 2023

September 2023

Data Understanding & Exploration

What information does our data present? How can we use that to develop our models?

Data Preparation

Initial data cleaning and preprocessing (missing null values, quality assurance).

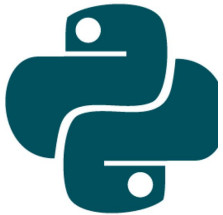
October 2023

November – December 2023

Model Development & Evaluation

Develop models and evaluate for accuracy.

Resources We Leveraged

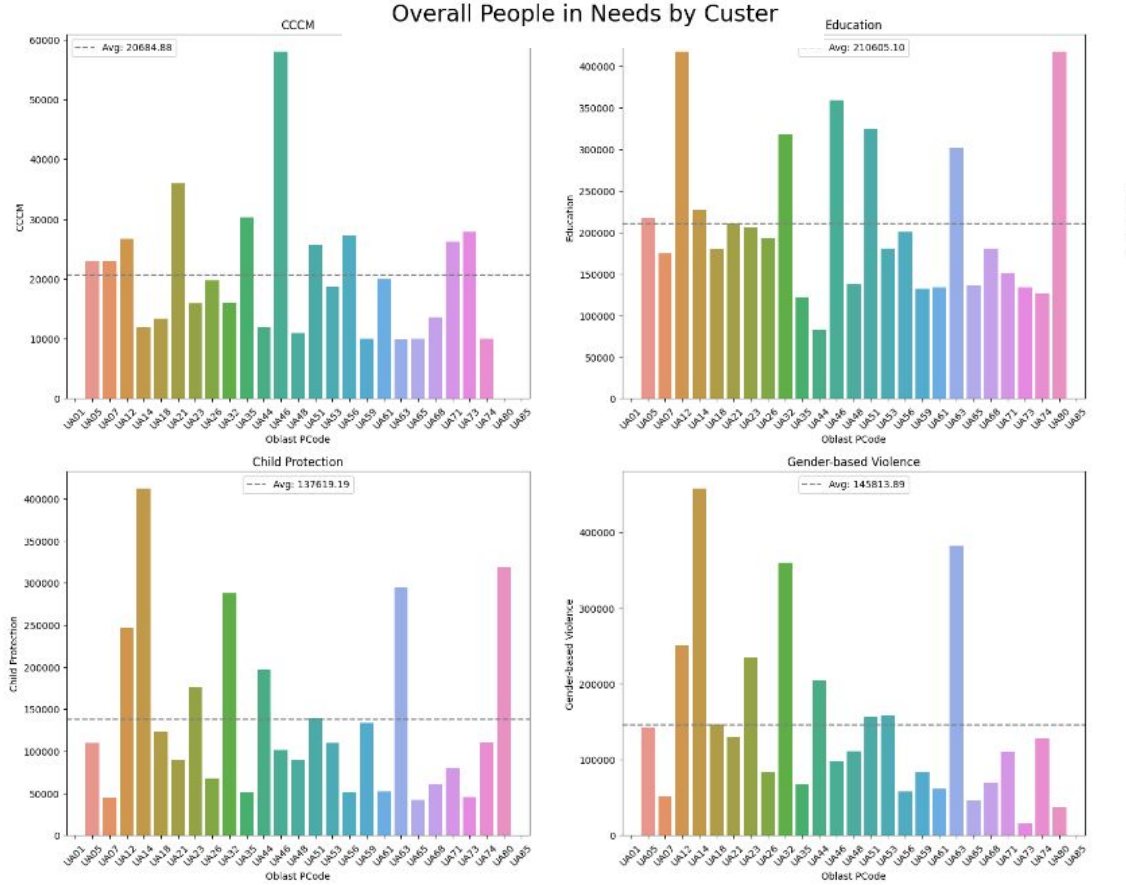


Data Understanding & Exploration

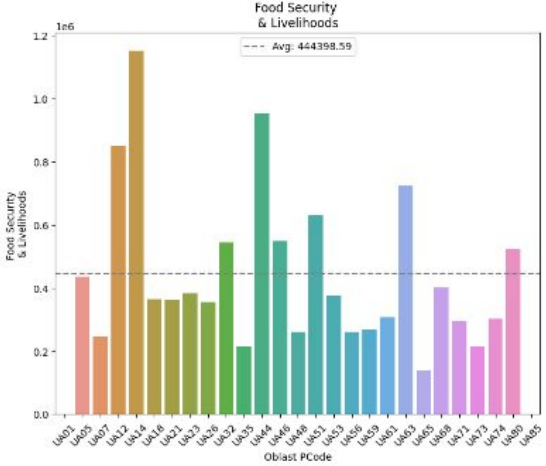
Dataset Overview

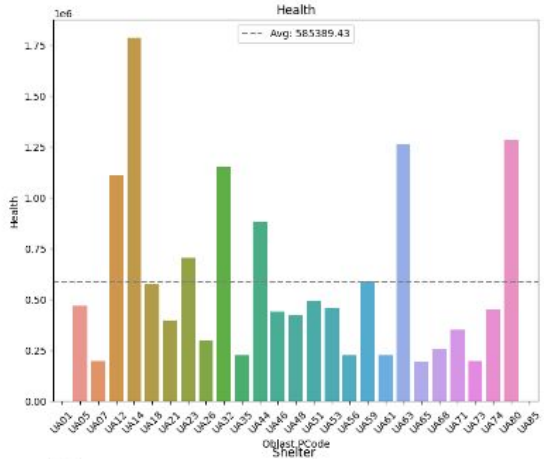
- Clustering
  - Ukraine: Humanitarian Needs Overview
- Forecasting Modeling
  - Ukraine Conflict Events: Civilian Targeting Events and Fatalities
- NLP
  - Russia vs Ukraine Tweets

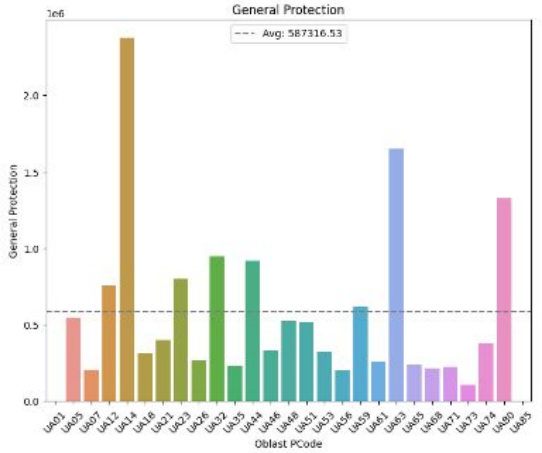
(Clustering) Ukraine: Humanitarian Needs Overview

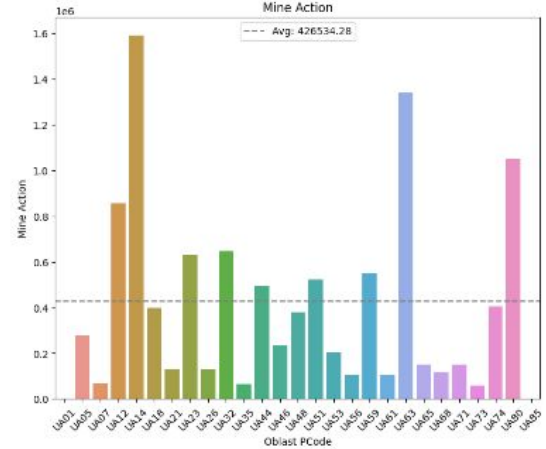


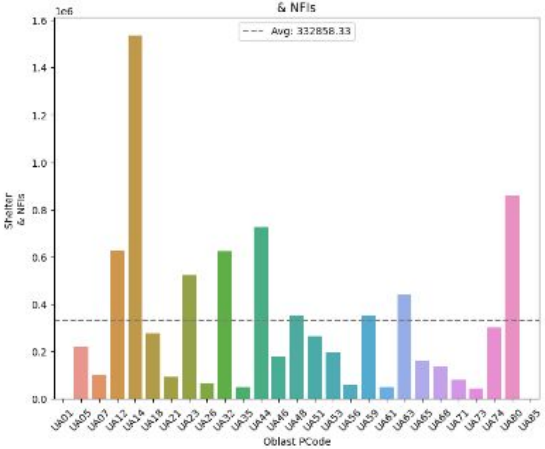
Overall People in Needs by Custer

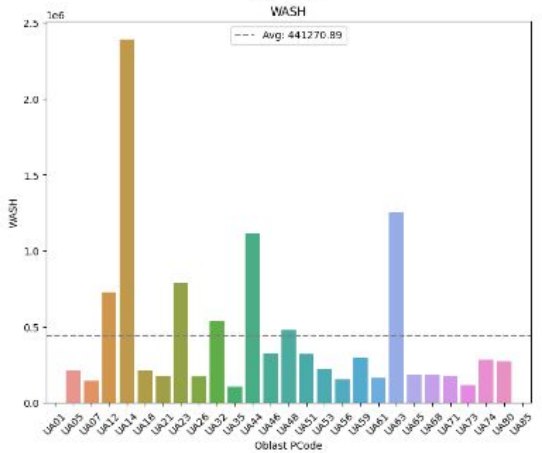




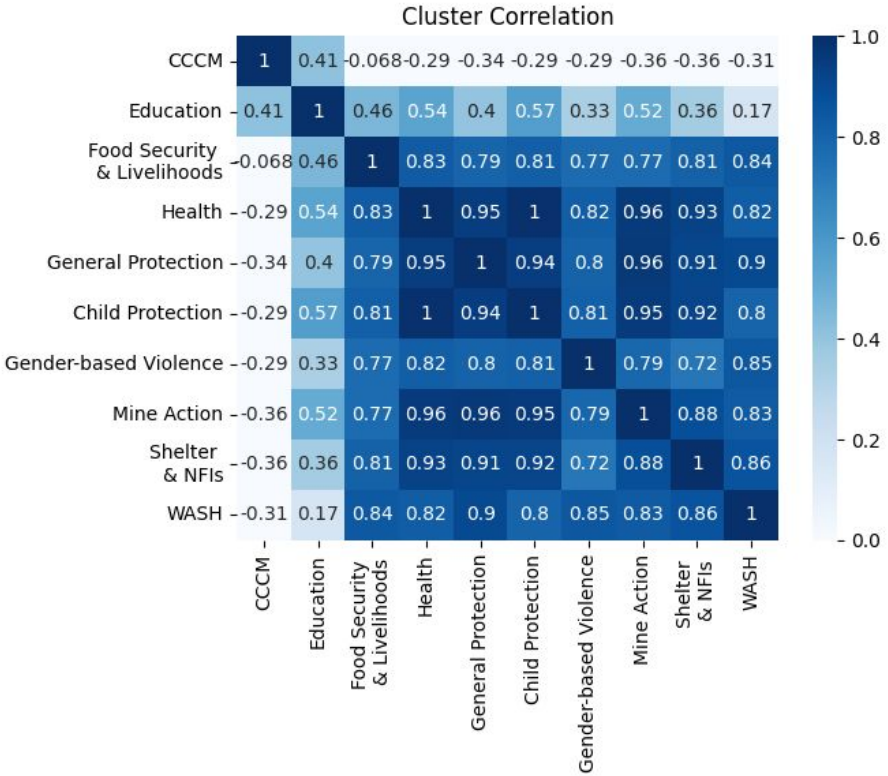


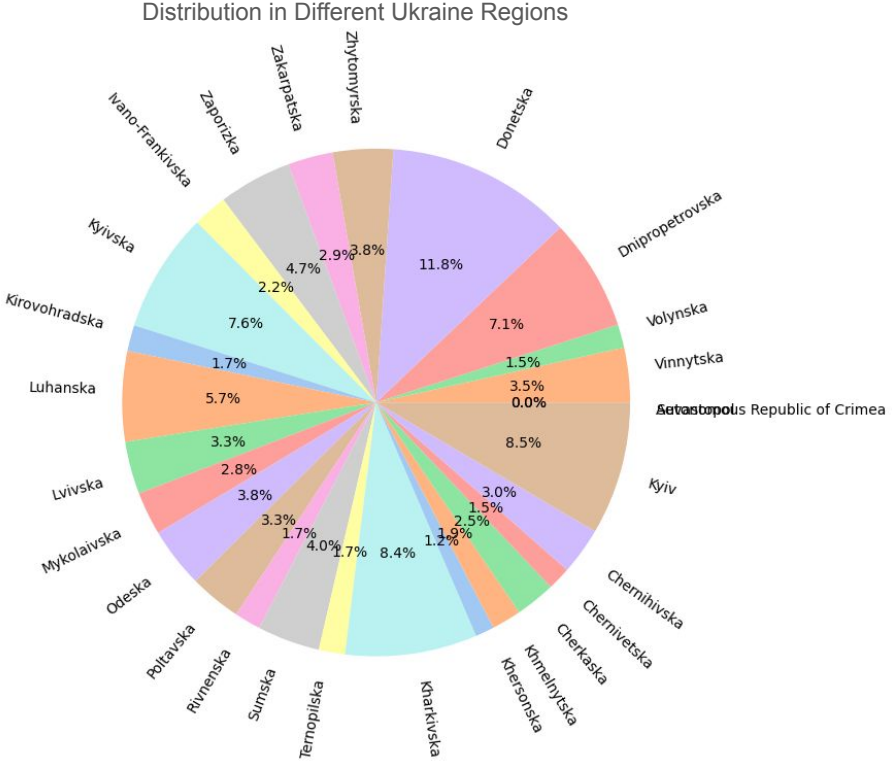




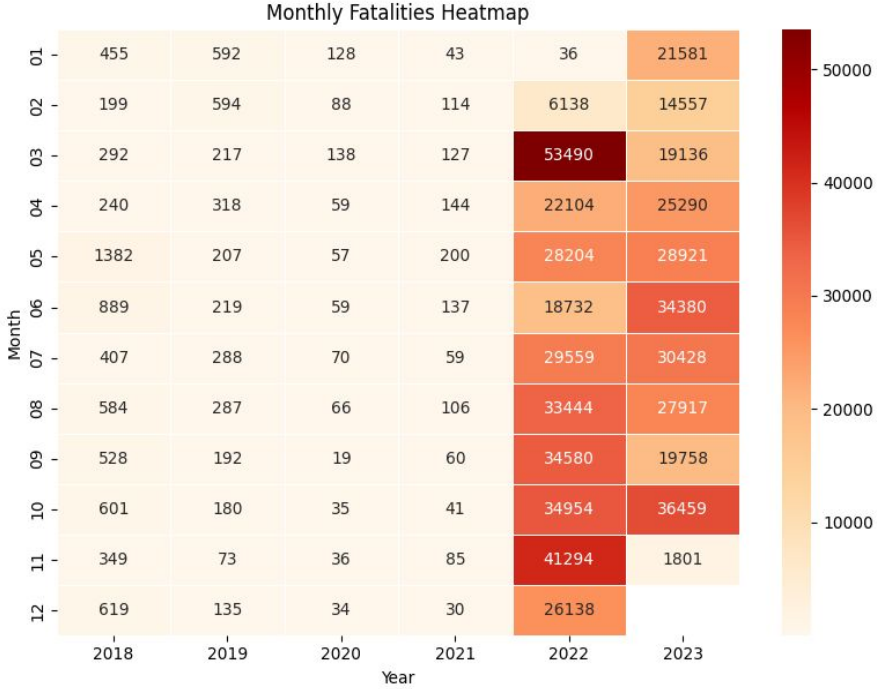


(Clustering) Ukraine: Humanitarian Needs Overvi

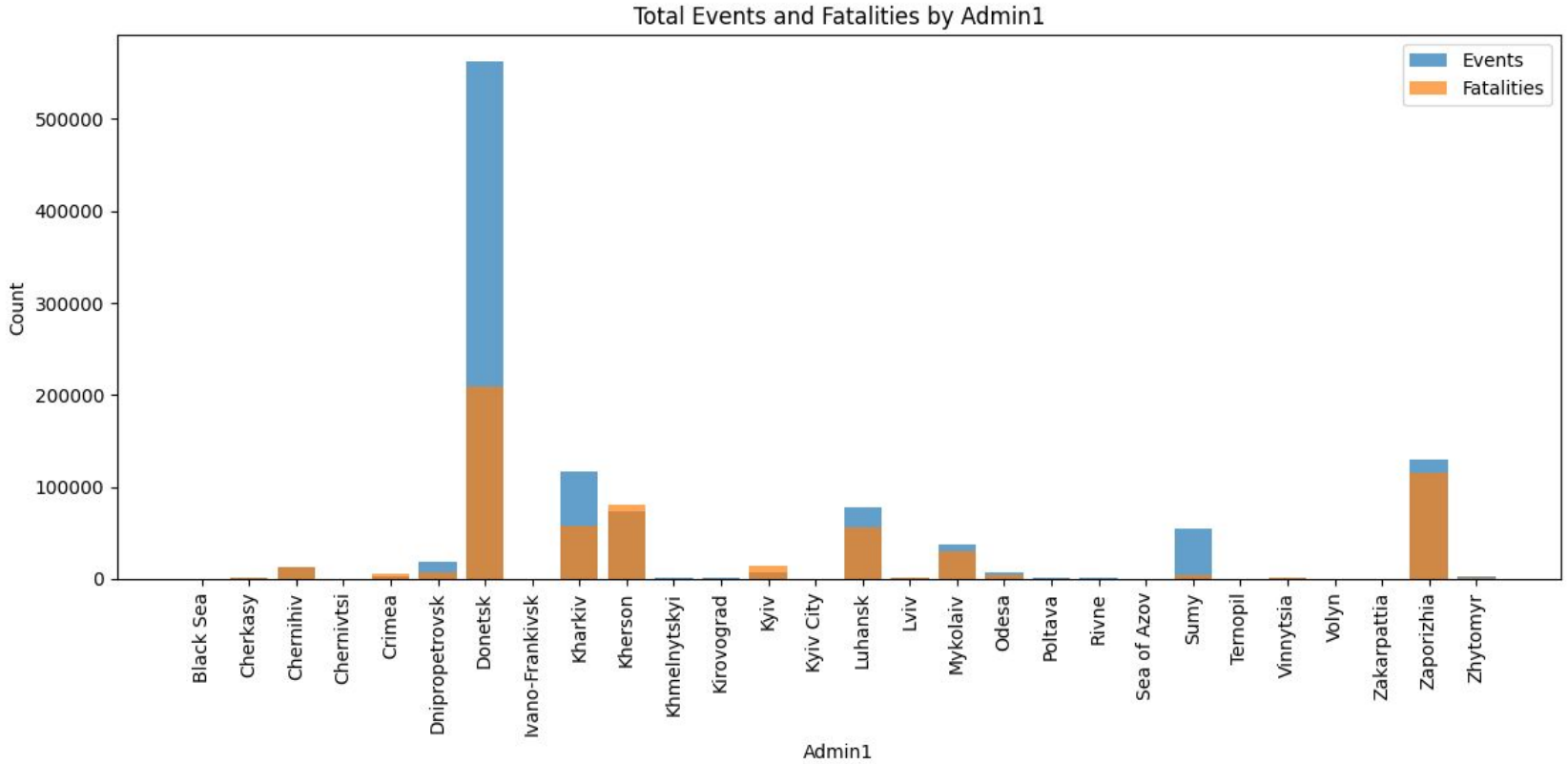




(Forecasting) Ukraine Conflict Events: Civilian Targeting Events and Fatalities



(Forecasting) Ukraine Conflict Events: Civilian Targeting Events and Fatalities



(NLP) Russia vs Ukraine Tweets Dataset

WordCloud: Most Frequently Used Words

Top 10 Important Words and Bi-grams

TF-IDF Analysis

russia: 2737.233130043287
ukrain: 1122.75408941269
war: 671.6390256146008
china: 641.354058280083
putin: 456.90775948465483
russian: 441.1370112125785
nato: 415.6137826321795
like: 392.2909854508427
countri: 375.2298084933223
would: 336.9751884595214

Bi-gram Analysis

russia china: 257.4012087352078
russia ukrain: 256.037926913401
ukrain russia: 229.1242423467909
war russia: 216.68802143045733
china russia: 180.6899634421418
support russia: 170.675425305153
russia invad: 152.36357472429117
war ukrain: 130.81548777504884
like russia: 120.77015898795523
ukrain war: 108.802386831676 39

Data Preparation & Preprocessing

Forecasting: Data Processing

Ukraine Conflict Events

- Data Cleaning
  - Removal of duplicate entries
- Feature Engineering:
  - Changing data types
  - Aggregation by Ukraine Cities
  - One-hot Encoding
  - CumSum - accumulating the number of events/fatalities by month
  - **(REVISED) Only retrieving data starting from Jan-2022 (Start of Ukraine-Russia War)**

Forecasting: Data Processing

Unrevised Approach

	Events	Fatalities	Admin1_Black Sea	Admin1_Cherkasy	Admin1_Chernihiv	Admin1_Chernivtsi	Admin1_Crimea	Admin1_Dnipropetrovsk	Admin1_Donetsk	Admin1_Ivano-Frankivsk	...	Admin1_Rivne	Admin1_Sea of Azov	Admi
5	13	1	0	0	0	1	1	2	2	0	...	0	0	
4	29	2	0	0	0	0	2	0	4	0	...	0	0	
8	56	8	0	0	0	1	1	1	3	0	...	0	0	

Revised Approach

	Events	Fatalities	Admin1_Black Sea	Admin1_Cherkasy	Admin1_Chernihiv	Admin1_Chernivtsi	Admin1_Crimea	Admin1_Dnipropetrovsk	Admin1_Donetsk	Admin1_Ivano-Frankivsk	...	Admin1_Rivne	Admin1_Sea of Azov	Admin1_Sumy	Admin1_Ternopil
53	684	138	0	0	0	0	0	0	0	0	...	0	0	0	0
52	764	293	0	0	4	0	0	1	5	0	...	0	0	4	0
56	1433	2538	0	0	5	0	0	4	7	0	...	1	0	4	0
49	1797	2960	0	0	3	0	2	5	7	0	...	0	0	4	0

NLP: Text Preprocessing Steps Applied to Tweets

- Converted all text to lowercase for consistency.
- Removed punctuation marks from the text.
- Broke down the text into individual words or tokens.
- Excluded common English stopwords to filter out unnecessary words.
- Reduced words to their base or root form.

Modeling & Evaluation

Algorithm Selection

- Forecasting Modeling
  - Autoregressive Integrated Moving Average (ARIMA)
  - Seasonal Autoregressive Integrated Moving Average (Seasonal ARIMA)
- Natural Language Processing
  - Keras: Sequential Model (Neural Network)

Forecasting Model Comparison

Autoregressive Integrated Moving Average (ARIMA)

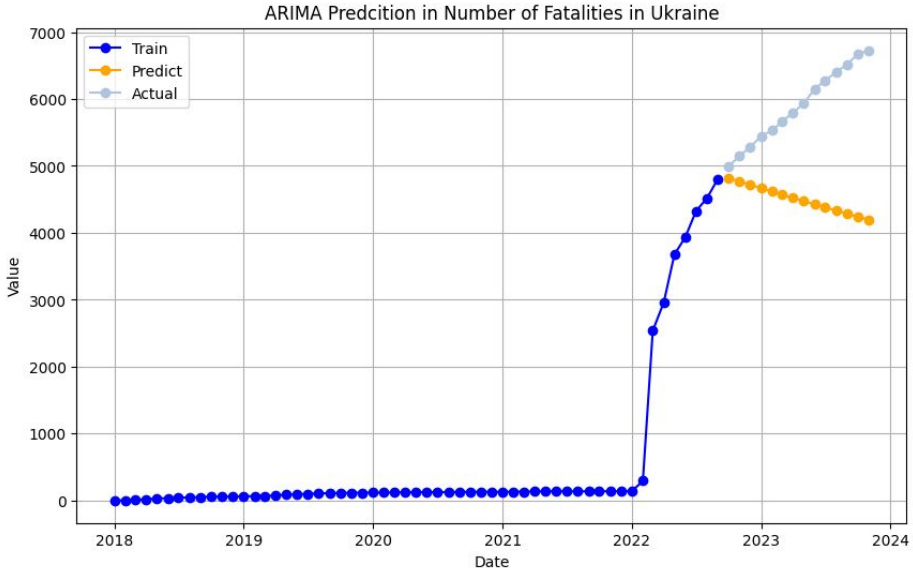
- **Autoregressive (AR):** Captures the relationship between the current observation and past values.
- **Integrated (I):** Transforms original data by taking differences between consecutive observations to make the time series data set more stationary.
- **Moving Average (MA):** Models the relationship between the current observation and residual errors from lagged (past) observations.
- Designed for time series data without clear seasonality.
- Good for short to medium-term forecasting.

Seasonal Autoregressive Integrated Moving Average (Seasonal ARIMA)

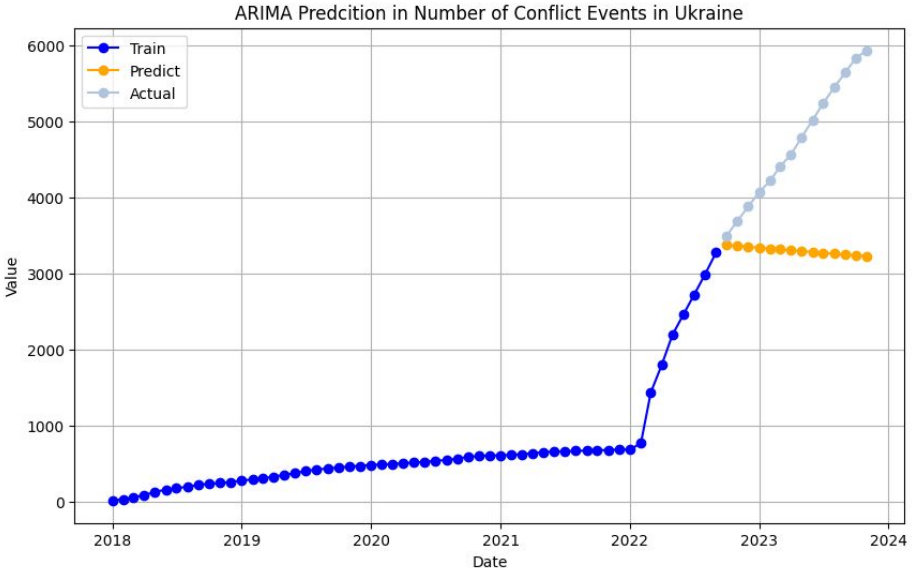
- Extension of the ARIMA model that incorporates seasonality into the forecasting process.
- Includes additional AR, I, and MA components at seasonal intervals.
- Designed to handle time series data that exhibits regular patterns or fluctuations at specific intervals, such as daily, monthly, or quarterly seasons.
- Useful for data with predictable seasonal patterns.
- Versatile for different time intervals.

Forecasting Model Training

- Autoregressive Integrated Moving Average (ARIMA)



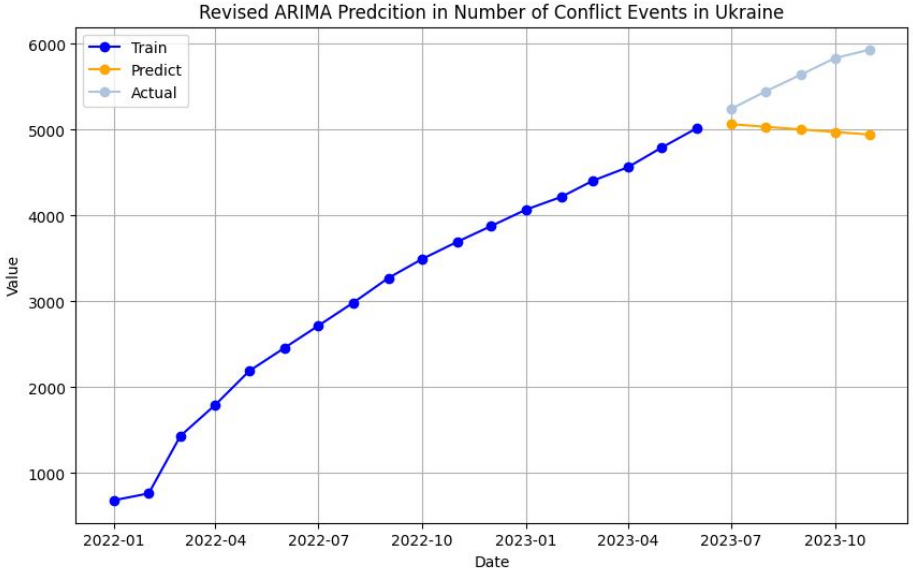
MAE, MSE, RMSE
[1472.8203978105544, 2776012.619508048, 1666.1370350328475]



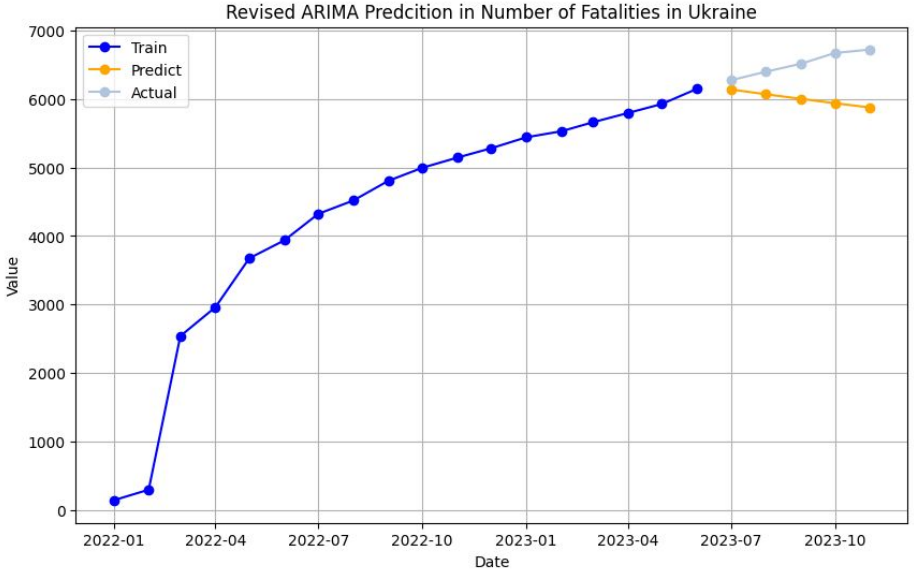
MAE, MSE, RMSE
[1518.0024305779307, 3043597.5386667727, 1744.5909373451339]

Forecasting Model Training

- Autoregressive Integrated Moving Average (ARIMA) Revised Approach



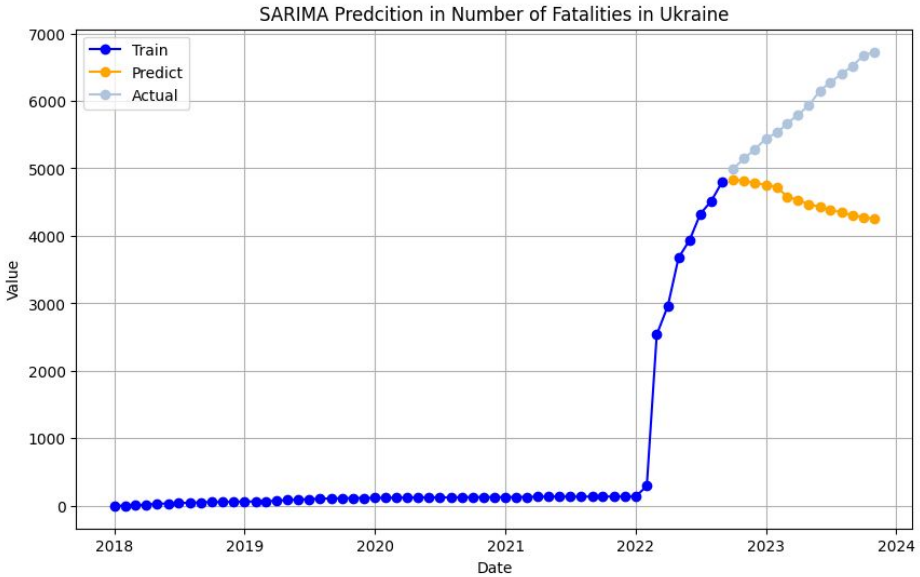
MAE, MSE, RMSE
[615.7107367894723, 465915.0404859345, 682.5796953366944]



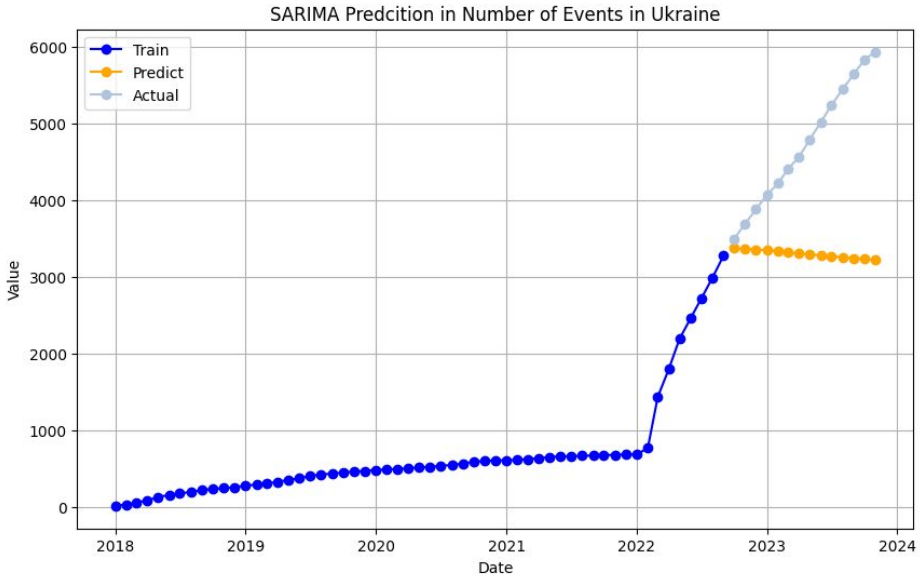
MAE, MSE, RMSE
[511.4782457181997, 328904.04245995387, 573.5015627354069]

Forecasting Model Training

- Seasonal Autoregressive Integrated Moving Average (Seasonal ARIMA)



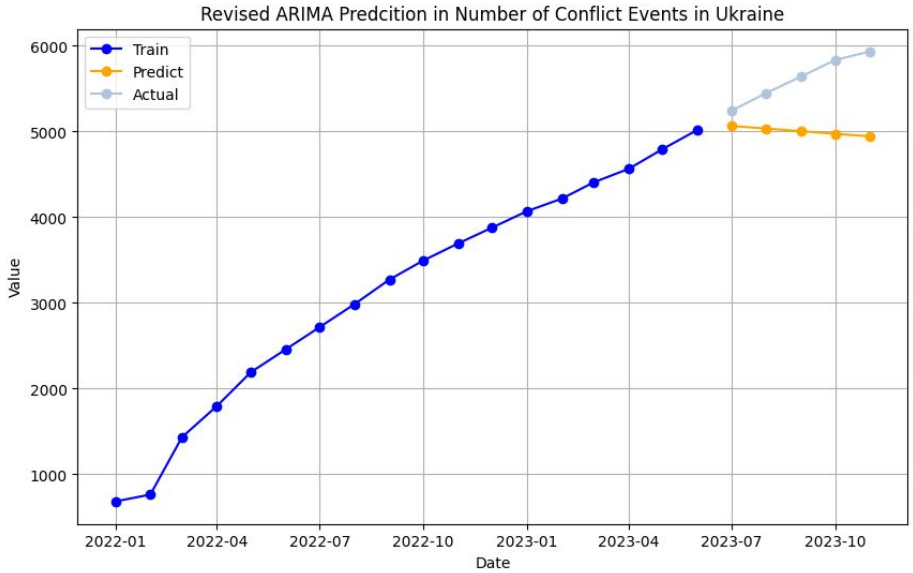
MAE, MSE, RMSE
[1436.0637634884392, 2677768.8422412076, 1636.3889642261731]



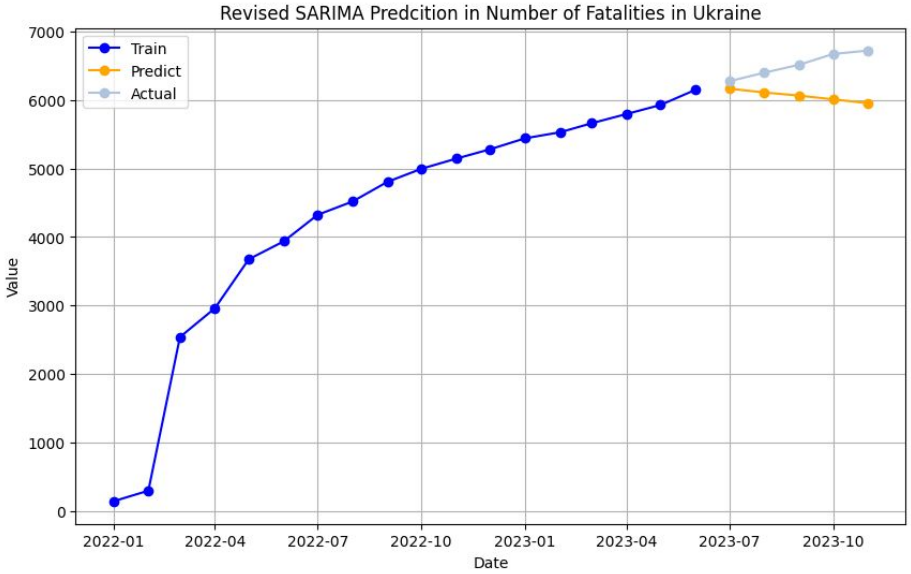
MAE, MSE, RMSE
[1517.3988514311038, 3048558.005011501, 1746.0120288851108]

Forecasting Model Training

- Seasonal Autoregressive Integrated Moving Average (Seasonal ARIMA) Revised Approach



MAE, MSE, RMSE
[301.0246236895722, 112718.22343999229, 335.73534731986786]



MAE, MSE, RMSE
[455.9732150878497, 265903.00236909126, 515.657834585194]

Keras: Sequential Model

The Keras library provides a high-level interface for building and training NN models, with a focus on simplicity and ease of use that allows you to easily organize the layers in a neural network.

Common Features of Keras Models

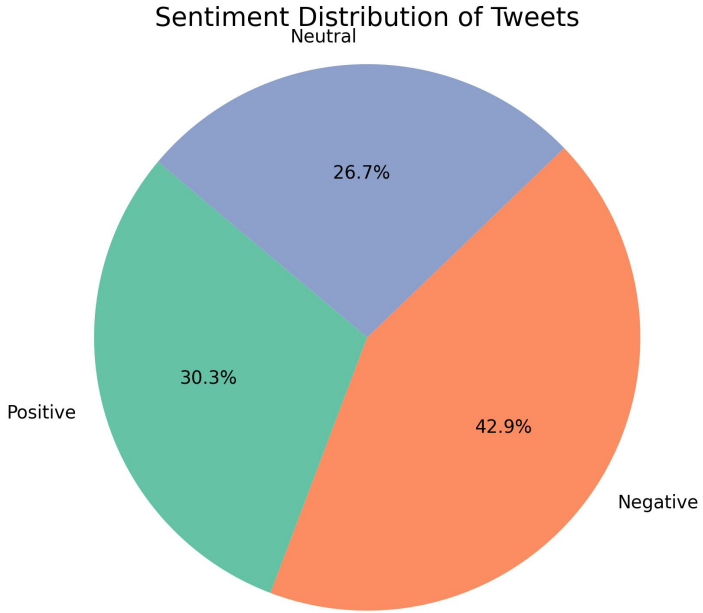
- **Compilation**: Before training a model, you need to configure the learning process by compiling the model with the compile method. It receives three arguments: an optimizer, a loss function, and a list of metrics.

- **Training**: After compilation, the model is trained with the fit method, which takes the input data, the target data, and the number of epochs to train, along with other optional parameters.

- **Evaluation and Prediction**: Models can be evaluated with the evaluate method and can make predictions on new data using the predict method.

Initial Sentiment Analysis Using SIA

- Negative: 42.9%
- Positive: 30.3%
- Neutral: 26.7%

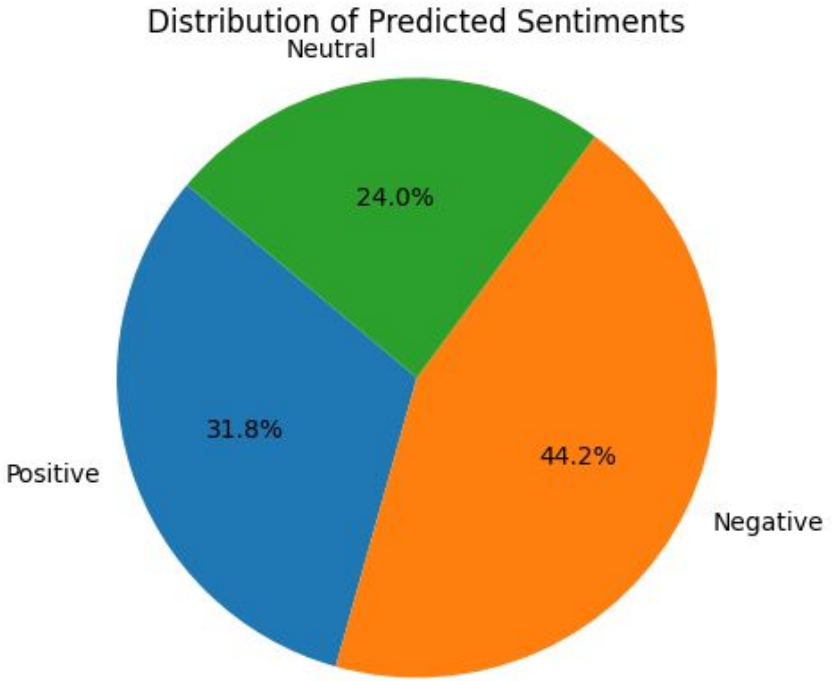


NLP Model Results

- Keras

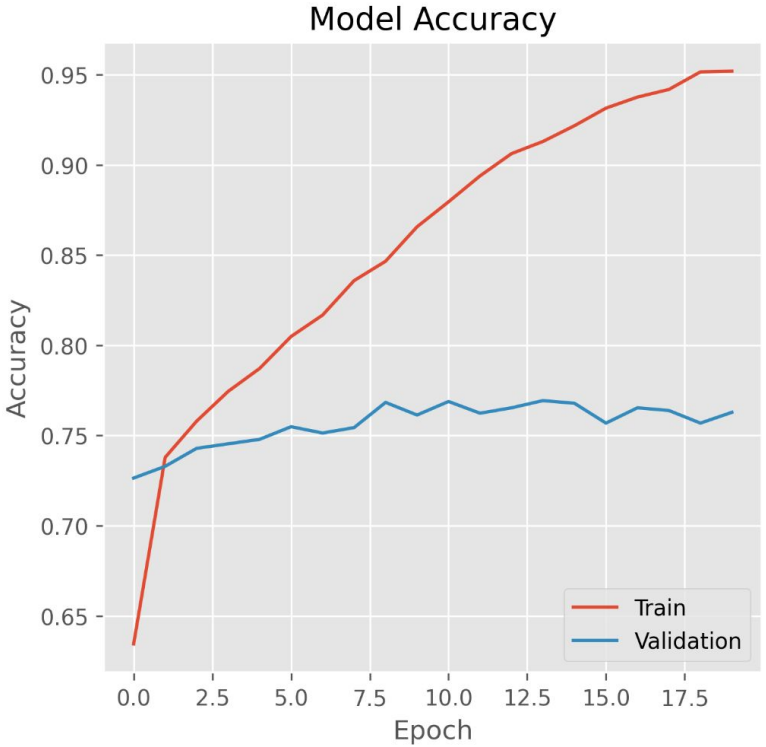
Predicted Sentiments:

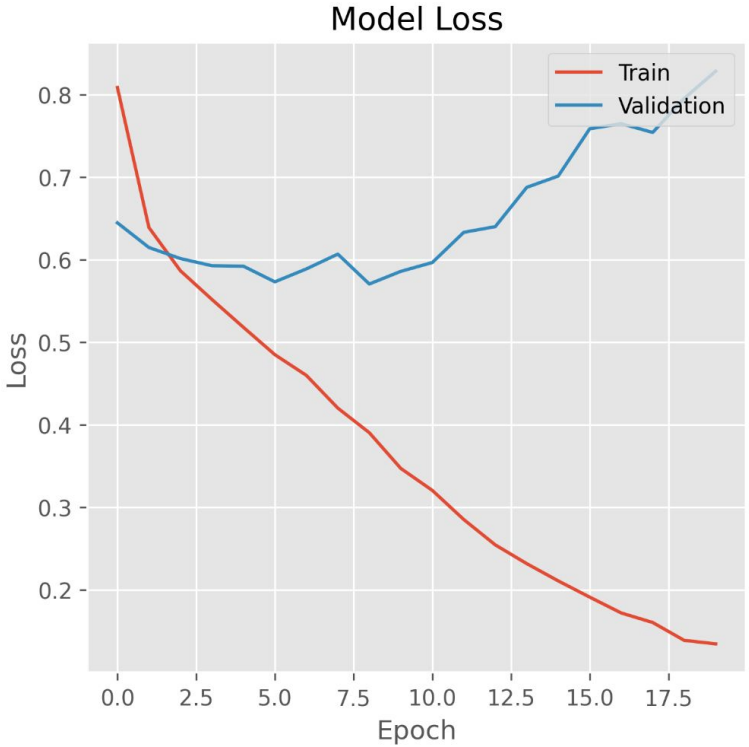
- Negative: 44.2%
- Positive: 31.8%
- Neutral: 24.0%



NLP Model Evaluation

- Keras - ATTEMPT 1





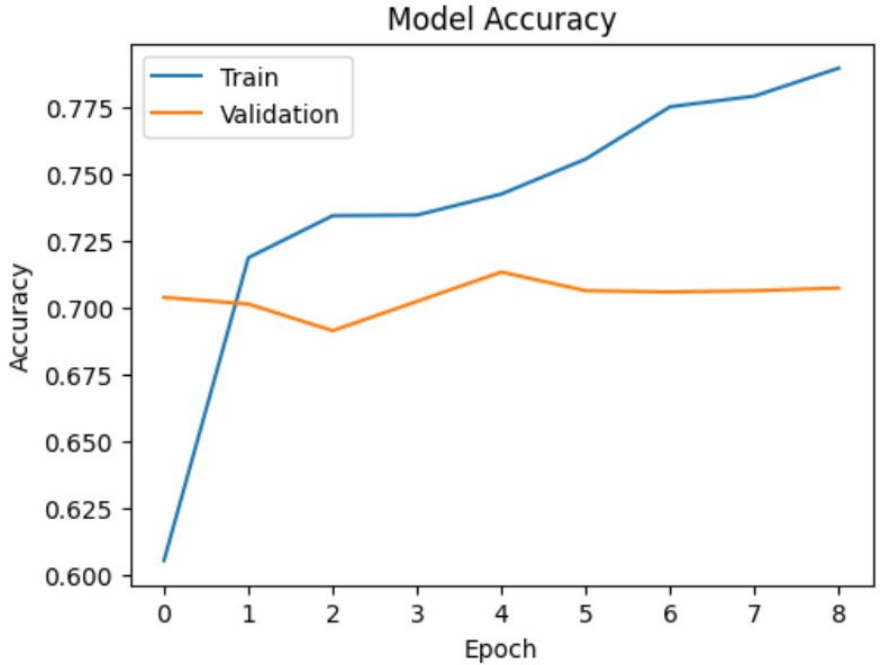
NLP Model Evaluation

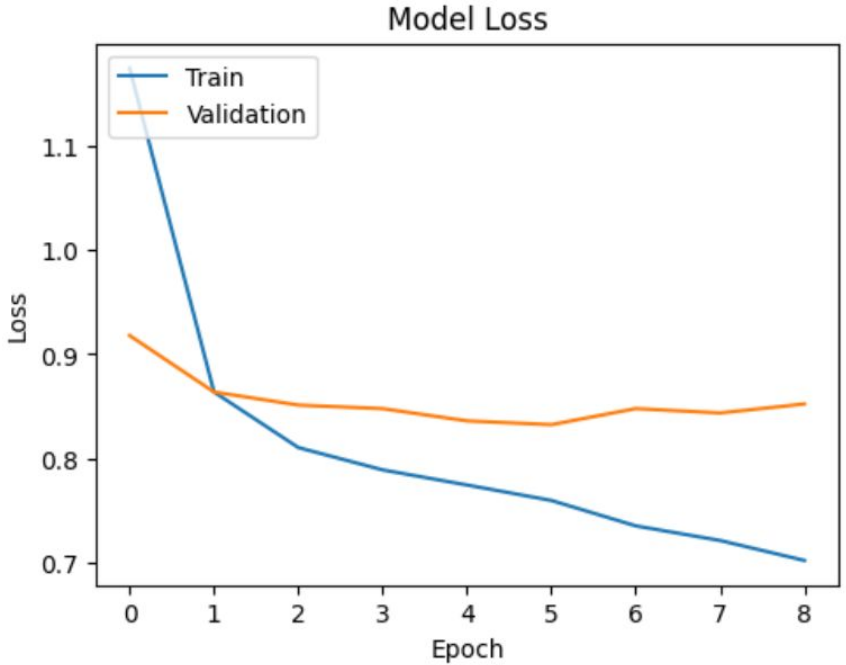
Keras - Improvements

- **Regularization**: reduces overfitting by adding a penalty to the loss function for large weights, encouraging the model to learn simpler patterns.
- **Early Stopping**: halts the training process when validation loss stops improving for a given number of epochs. This prevents the model from continuing to learn from the training data to the point where it starts to overfit
- **Learning Rate Scheduling**: Learning rate scheduling changes the learning rate during training, reducing it gradually. Helps prevent overfitting by giving the model more time to fine-tune it’s weights.

NLP Model Evaluation

- Keras - ATTEMPT 2





Final Thoughts

Insights and Key Findings

- Clustering
  - Data visualization of distribution between clusters shows different level of people in needs across different regions in Ukraine. Especially, UA12 - Dnipropetrovska has the highest needs across different clusters, which it might need further aids in the future.
- Forecasting
  - In both the Seasonal ARIMA and the ARIMA forecasting models, we found that the model's predicted number of fatalities and conflict events was less than the actual value.
  - This implies that we underestimate the effect of the war on the number of people in need.
- NLP Sentiment Analysis
  - Initial sentiment analysis performed on the tweets dataset shows a distribution of about 30% positive and 43% negative. However, the keras predictive model shows 32% positive and 44% negative.

Potential Next Steps

- Use more granular data (daily or weekly) for forecasting model for more accurate performance.
- Find dataset of tweets over longer period of time (pre-conflict to post-conflict) for NLP to perform temporal analysis on sentiment changes toward Russia.
- Incorporate models and visualizations into an interactive Tableau story.

THANK YOU!

Questions?