Sintia Stabel
sstabel@syr.edu
IST – 736 Text Mining HW5
Sklearn TF-IDF vectorizer
Sklearn Multinomial Naïve Bayes
Sklearn Random Forest

Aug 8th, 2023

Severe Weather Event Classification: Unveiling Insights from Recent News Articles

Introduction:

Severe weather events, including wildfires, droughts, and heatwaves, have become increasingly prevalent in recent years, posing significant challenges to communities worldwide. In this era of abundant information, recent news articles have emerged as a valuable and readily accessible source of data to understand the impact and implications of these extreme weather phenomena. These articles provide real-time accounts of unfolding situations, expert analysis of damages, and insights into response efforts, making them an invaluable resource for decision-makers and the public alike.

During severe weather events, emergency management authorities require up-to-date and comprehensive information to assess the situation and allocate resources effectively. By categorizing recent news articles, these authorities can gain valuable insights to prioritize response efforts, implement evacuation plans, and deploy resources where they are most needed. In addition, climate researchers and scientists can analyze valuable data on the frequency, intensity, and geographical distribution of severe weather events.

Governments, policymakers, as well as non-governmental organizations (NGOs) working on disaster relief and response can leverage categorized news articles to identify regions that require urgent assistance and provide targeted aid to affected communities. They need accurate and real-time information to develop effective policies and strategies to address the growing challenges posed by wildfires, droughts, and heatwaves. By accurately classifying news articles into distinct groups of "wildfire," "drought," and "heatwave," decision-makers, emergency responders, and the public can obtain timely information and gain a deeper understanding of the challenges posed by these extreme weather phenomena. Categorized news articles offer a snapshot of the situation on the ground, enabling expedited data-driven policy decisions.

Analysis

Data Preparation and Cleaning

The dataset was obtained from NewsAPI, a simple rest API that returns search results for current and historic news articles published by over 80 thousand sources worldwide. A request containing the following topics was submitted: “wildfire”, “heatwaves” and “droughts”. The results of the request were then transformed and saved from a JSON object into saved in a csv file with 300 rows of data and 5 columns. The column names and content descriptions are as follows:

LABEL: contains the topic of the article: wildfire, heatwaves, or droughts.

Date: date on which the article was published.

Title: contains the title of the article.

Headline: contains the articles' headline, which is the equivalent of a short snippet of what the article is about. Figure 1 shows a sample of the resulting dataset.

Figure 1 – Raw data in JSON format from NewsAPI


```
<Response [200]>
{'status': 'ok', 'totalResults': 3381, 'articles': [{'source': {'id': None, 'name': 'Lifehacker.com'}, 'author': 'Eliza
beth Yuko', 'title': 'Use This Phone Number to Find a Cooling Center Near You', 'description': 'We're not even a full m
onth into summer, and much of the United States has already dealt with record-setting heat—not to mention the intermitt
ent wildfire smoke from Canada. Between climate change and El Niño, this type of weather is expected to get worse du
r…', 'url': 'https://lifehacker.com/use-this-phone-number-to-find-a-cooling-center-near-you-1850641814', 'urlToImage':
'https://i.kinja-img.com/gawker-media/image/upload/c_fill,f_auto,fl_progressive,g_center,h_675,pg_1,q_80,w_1200/c3be802
7600a57f72bc8352f391747a0.jpg', 'publishedAt': '2023-07-15T15:00:00Z', 'content': 'Were not even a full month into summ
er, and much of the United States has already dealt with record-setting heatnot to mention the intermittent wildfire sm
oke from Canada. Between climate change and … [+1728 chars]'}, {'source': {'id': 'bbc-news', 'name': 'BBC News'}, 'auth
or': None, 'title': 'Italy: Drone spots suspected wildfire arsonist', 'description': 'A man is detained in Italy after
he was seen at the scene of a wildfire in Calabria.', 'url': 'https://www.bbc.co.uk/news/av/world-europe-66343164', 'ur
lToImage': 'https://ichef.bbci.co.uk/news/1024/branded_news/160AF/production/_130578209_p0g3kd5k.jpg', 'publishedAt': '
2023-07-28T17:21:56Z', 'content': 'A man has been detained after he was spotted by a drone near a recently started wild
```


Figure 2 – First 10 rows of csv file: severe_weather.csv

	LABEL	Date	Source	Title	Headline
0	wildfire	7/15/2023	Lifehacker.com	Use This Phone Number to Find a Cooling Center...	even full month into summer much United States...
1	wildfire	7/28/2023	BBC News	Italy Drone spots suspected wildfire arsonist	detained Italy after seen scene wildfire Calabria
2	wildfire	7/30/2023	BBC News	Canada wildfire Firefighter dies tackling Brit...	evacuation order place towns near border fires...
3	wildfire	7/18/2023	Google News	Wildfire rages near Athens Reuters	Wildfire rages near Athens Reuters
4	wildfire	7/23/2023	Google News	Jet TUI cancel flights to Rhodes because of wi...	cancel flights Rhodes because wildfire Reuters

Data cleaning and preparation involved performing text preprocessing on the Headline column to ensure the data is suitable for classification analysis. To analyze the text data, computerized processing methods were used to convert text data into vectorized data, from which computer algorithms can extract word frequencies and conduct further analysis. Prior to vectorization, however, the contents from the variable "Headlines" were preprocessed with the following actions to produce more accurate results during modeling:

- Special Characters and Punctuation: regular expressions (regex) were used to remove any characters that are not alphanumeric or spaces from the text, including numerical digits.
- Convert Text to Lowercase: all text was converted to lowercase to ensure uniformity in text representation. This step was done to avoid any issues related to case sensitivity during analysis.
- Tokenization: The text was tokenized using NLTK's word_tokenize function. Tokenization breaks the text into individual words, which are referred to as tokens.
- Stopwords Removal: Stopwords are common words that do not add much meaning to the text (e.g., 'the', 'and', 'is'). The function removes stopwords using NLTK's set of English stopwords, retaining only meaningful words in the text.
- Removal of Nan Values: Missing values can hinder a model's ability to learn from the data and make accurate predictions. The rows containing nan values were dropped as in this case there was only one occurrence in the entire dataset.

Figure 3 – New Column with preprocessed and clean data: "df_processed"

	LABEL	Date	Source	Title	Headline	df_processed
295	Droughts	7/20/2023	Forbes	Rivers As Weapons Of War And Warnings Of Clima...	millennia armies have used rivers weapons Whil...	millennia armies used rivers weapons intention...
296	Droughts	7/21/2023	Science Daily	Climate science is catching up to climate chan...	Africa climate change impacts experienced extr...	africa climate change impacts experienced extr...
297	Droughts	7/15/2023	Home.blog	Loops and Arcs	Here tools been using lately better understand...	tools using lately better understand functiona...
298	Droughts	7/24/2023	Phys.Org	Algeria fires fanned by winds extreme heat kill	Wildfires raging across Algeria during blister...	wildfires raging across algeria blistering hea...
299	Droughts	8/3/2023	Phys.Org	Geostationary satellite reveals widespread mid...	western particularly Southwest experienced not...	western particularly southwest experienced not...

Data Exploration

Having completed the data cleaning and preprocessing, the severe_weather dataset unveils deeper insights through engaging WordCloud visualizations. A WordCloud offers a concise snapshot of the dataset's pivotal words, highlighting their frequency. The size of a word corresponds to its prevalence within the dataset. The accompanying Figures 4 to 6 present these WordClouds, each pertaining to the distinct topics or labels found within the dataset.

For instance, the wildfire WordCloud accentuates essential terms such as "wildfire," "fire," "firefighter," "Canada," "evacuation," and "killed." In the heatwave WordCloud, prominent words include "temperature," "heatwave," "extreme," "hottest," "record," and "wildfire." Similarly, the drought WordCloud underscores significant words like "climate," "change," "drought," "global," "water," and "wildfire."

Figure 4 – WordCloud of wildfire headlines data

Figure 5 - WordCloud of heatwave headlines data

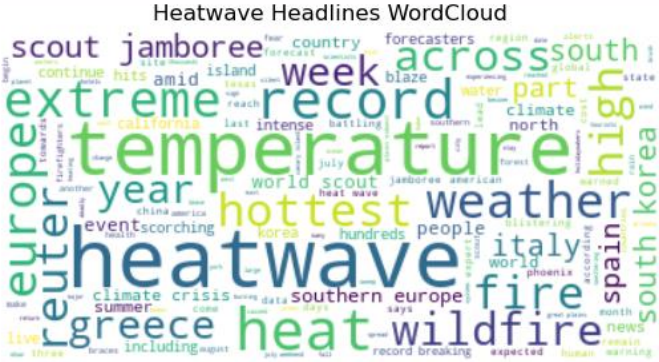
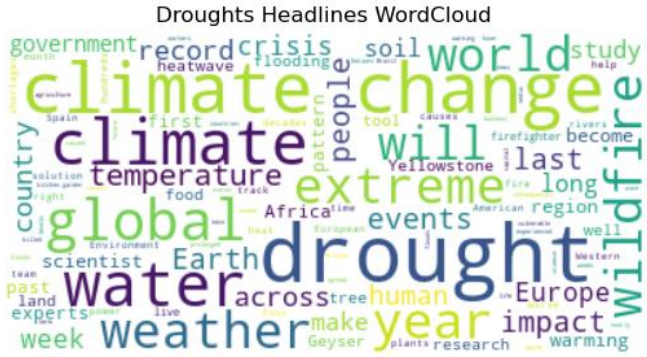


Figure 6 – WordCoud droughts headlines data



In addition to the article snippets and titles corresponding to each topic or label, the dataset also encompasses the sources from which the articles originated. This information provides valuable context. The image depicted below illustrates the predominant sources for each topic.

Observing the stacked bar chart reveals prominent patterns. For articles concerning wildfire and heatwave, Google News and Fort Worth Star-Telegram emerge as the predominant sources, with BBC News trailing closely. Conversely, drought-related articles predominantly emanate from Al Jazeera English and Fort Worth Star-Telegram, with Forbes ranking as a secondary source. This presentation underscores the influencend prominence of specific news outlets within each distinct weather-related category.

Figure 7 - Top news sources distribution by weather event

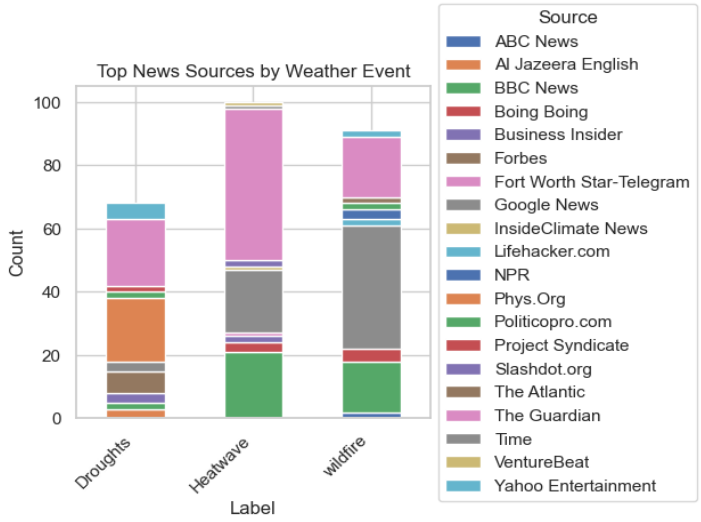
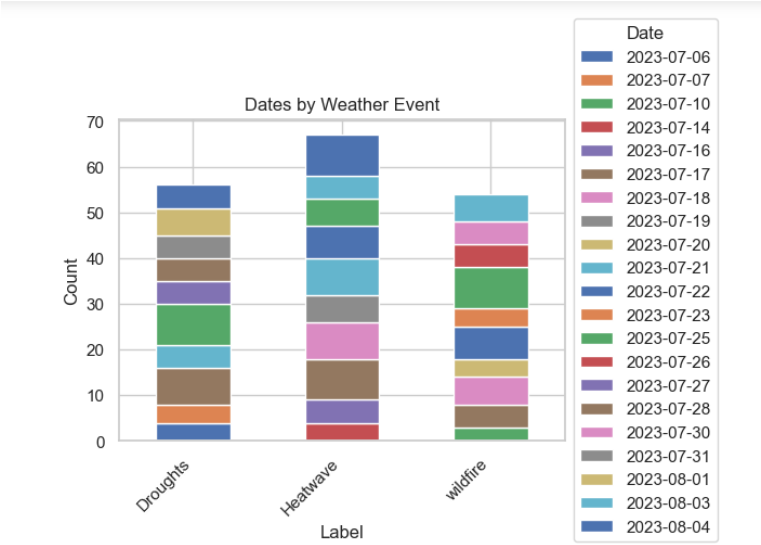


Figure 7 below visually represents the publication dates of the articles. These dates are arranged in descending order of frequency, showcasing the most frequently occurring publication dates. Specifically, the dates with the highest frequencies are July 10th and 6th, followed closely by July 17th and 18th.

A vital consideration is that the date range for the dataset spans from July 6th, 2023 to August 5th, 2023. It is noteworthy that the number of articles referencing these events within this timeframe does not provide a comprehensive global count. This limitation arises from the dataset's dependency on the features provided by the News API's Rest API, which serves as the primary data source for this study.

Figure 8 – Top article dates distributions by event type



Leveraging the spaCy library within NLTK, a systematic approach was employed to extract locations mentioned within the article titles. This endeavor aimed to furnish a comprehensive glimpse into the geographic regions most frequently associated with the transpiring weather events.

Figure 9 – Weather event count frequency by location

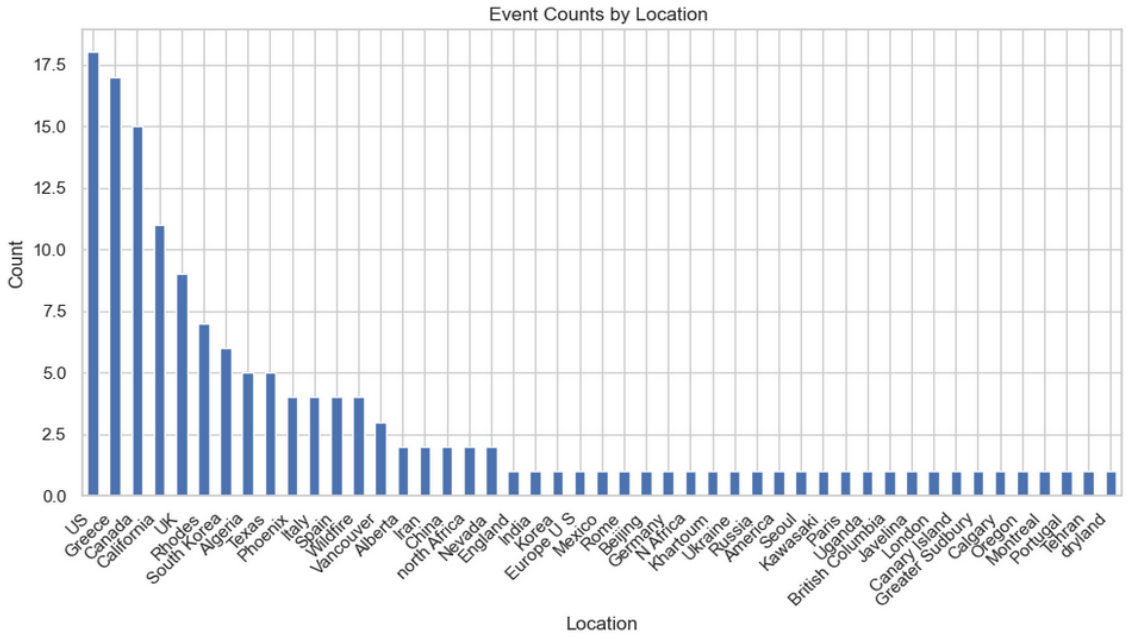
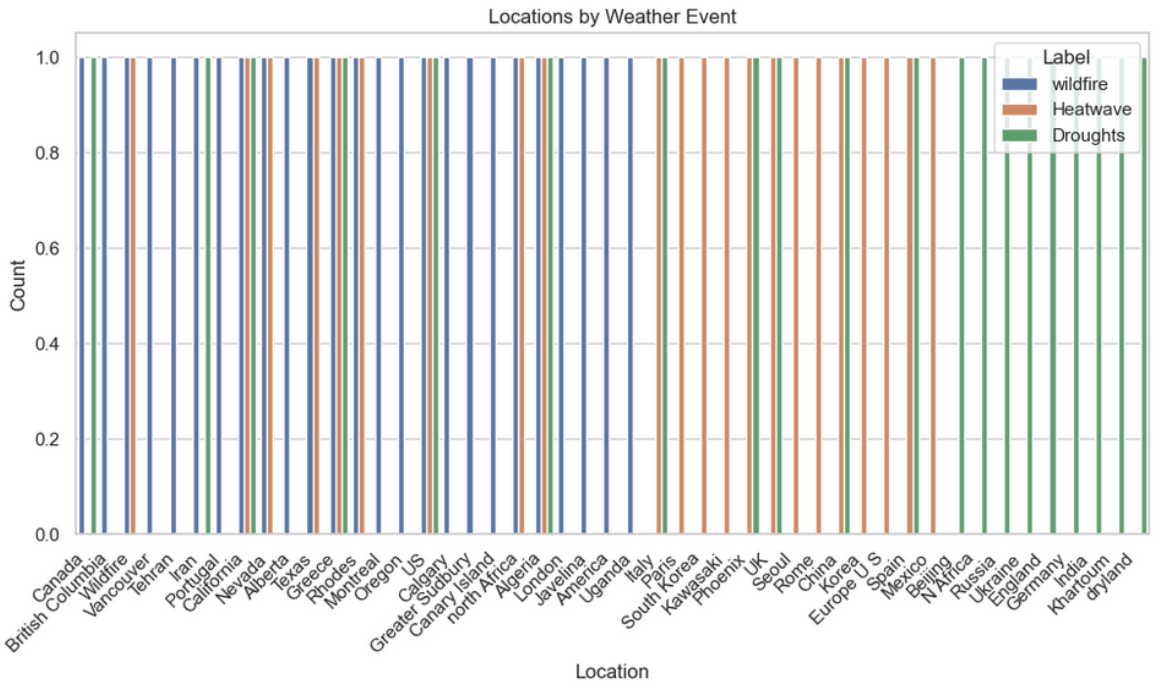


Figure 10 shows a higher frequency of wildfires referencing Canada and US states and territories. While heatwaves and droughts articles are mostly referencing European and Asian countries. Northern Africa and Uganda are more frequently associated with wildfires in this dataset.

Figure 10 – Location frequency by event type



Models and Methods

TF-IDF Vectorizer

For the analysis of the severe_weather dataset, the TF-IDF vectorization method was employed. TF-IDF, which stands for "Term Frequency-Inverse Document Frequency," is a technique used to transform textual data into numerical features suitable for machine learning. This approach plays a crucial role in converting text data into a format that can be utilized by machine learning algorithms. It works by calculating a score for each term in the text, indicating its significance in a specific document relative to the entire dataset.

The term frequency (TF) component of TF-IDF measures how frequently a term appears in a document, giving insight into the importance of that term within that document. Conversely, the inverse document frequency (IDF) component quantifies the rarity of a term across all documents. Terms that are rare and appear in only a few documents are assigned higher IDF scores, highlighting their unique nature and potential importance. By combining the TF and IDF scores, the TF-IDF vectorizer generates a numerical representation for each document, where term importance is weighted based on its frequency within the document and rarity across the dataset.

Text Preprocessing: Lemmatization and Stemming

In addition to employing TF-IDF vectorization, the text data underwent preprocessing steps known as lemmatization and stemming. These techniques were applied to enhance the quality and consistency of the textual content before feeding it into the machine learning models.

Lemmatization: Lemmatization involves reducing words to their base or dictionary forms to normalize variations of words. For instance, "running" and "ran" would be lemmatized to "run." This process helps ensure that different inflections of a word are treated as the same term, leading to better feature representation and improved model performance.

Stemming: Stemming, on the other hand, involves removing prefixes and suffixes from words to reduce them to their root forms. For example, "running" and "runner" would both be stemmed to "run." While stemming is more aggressive than lemmatization, it can sometimes result in words that are not actual words, but the technique is useful in simplifying words to their core meanings.

Model 1: Multinomial Naive Bayes (3 Cross Validation)

The Multinomial Naive Bayes (MNB) model is a specific variant of the Naive Bayes algorithm that is commonly used for text classification tasks, such as sentiment analysis or document categorization. It is particularly suitable for problems with discrete features, like word frequencies in text data. The initial model employed was a basic Multinomial Naive Bayes classifier with 3-fold cross-validation (3CV). This model utilized the TF-IDF vectorized representations of the documents to predict the categories "wildfire," "heatwaves," and "droughts." The Multinomial Naive Bayes algorithm makes predictions based on the probabilities of each class, utilizing the frequency of terms in the document.

Model 2: Naive Bayes + Lemmatization

Building upon the first model, the second model incorporated lemmatization as part of the text preprocessing. Lemmatization aimed to normalize words, reducing variations, and capturing the core meaning of terms. This model then underwent 3-fold cross-validation to assess its performance in classifying the document categories.

Model 3: Naive Bayes + Stemming

In the third model, stemming was applied to the text data as part of the preprocessing pipeline. Stemming helped to further simplify words by removing prefixes and suffixes. The model, similar to the first and second models, was evaluated using 3-fold cross-validation to determine its classification accuracy.

Model 4: Random Forest + Lemmatization

The final model introduced a different algorithm: Random Forest. This model utilized TF-IDF vectorization along with lemmatized text data. Random Forest is an ensemble learning algorithm that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) of the individual trees. It is particularly effective in handling complex relationships and avoiding overfitting.

At the core of the Random Forest algorithm are individual decision trees. Decision trees are hierarchical structures that break down a dataset into smaller and more manageable subsets by asking a series of binary questions based on feature attributes. These questions are designed to effectively split the data into different classes, ultimately leading to a classification decision at the tree's leaves.

The decision tree works by recursively splitting the data into subsets, with each split being determined by a feature and a threshold. The algorithm aims to find the best feature and threshold that results in the purest subsets, where samples within each subset predominantly belong to a single class (label). The process continues until a predefined stopping criterion is met, such as a maximum depth or a minimum number of samples in a leaf node.

While decision trees have the potential to model complex relationships in the data, they are prone to overfitting, capturing noise in the data and leading to poor generalization to unseen data. Random Forest builds on the strength of decision trees while addressing their limitations. Instead of relying on a single decision tree,

Random Forest creates an ensemble of multiple decision trees, each trained on a different subset of the data and with some randomness introduced.

In summary, these models and methods were systematically explored to gauge their performance in classifying documents related to "wildfire," "heatwaves," and "droughts." The incorporation of lemmatization and stemming aimed to enhance the quality of the input data, and the utilization of different models aimed to capture diverse patterns and relationships in the text. Cross-validation was employed to Naïve Bayes models to validate the models' performance and ensure their robustness on unseen data.

Results

TF-IDF Vectorizer Results

Below, in Figures 11 and 12, a portrayal of the initial 30 features or words from the processed vocabulary, derived through the TF-IDF vectorizer applied to the reviews data, is presented. In Figure 12, a visual representation of the first 10 rows and columns showcases the TF-IDF scores associated with each feature or word, subsequent to the meticulous cleansing and pre-processing steps.

It is crucial to acknowledge that the complete ensemble of features or vectors is considerably expansive, encompassing a substantial 299 rows and 2157 columns within the dataframe. Hence, the complete presentation of this extensive data is unfeasible. A closer examination discloses a marked discrepancy in the impact of stemming compared to lemmatization. For instance, terms like "abating" and "abnormal" undergo substantial reduction to "abat" and "abnorm" through stemming. On the other hand, the sample data depicted reveals minimal detectable change after lemmatization. This divergence underscores the fact that while stemming and lemmatization are both aimed at streamlining and simplifying the vocabulary, the outcomes they yield can markedly differ.

Figure 11 – First 30 severe weather data vocabulary words/features

```
['abating' 'ability' 'able' 'abnormally' 'academic' 'access' 'accom'
 'accordi' 'according' 'account' 'accounts' 'accused' 'acres' 'acropolis'
 'across' 'action' 'activist' 'adams' 'adapting' 'address' 'adriana'
 'advent' 'adverse' 'advisories' 'aerial' 'affect' 'affected' 'africa'
 'afternoon' 'agencies']
```

Figure 12 – Sample TF-IDF score assigned after vectorization of the headline data

	abating	ability	able	abnormally	academic	access	accom	accordi	\
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figure 13 – First 30 vocabulary features after Lemmatization and vectorization on headline data

```
['abating' 'ability' 'able' 'abnormally' 'about' 'above' 'academic'
 'access' 'accom' 'accordi' 'according' 'account' 'accused' 'acre'
 'acropolis' 'across' 'action' 'activist' 'adams' 'adapting' 'address'
 'adriana' 'advent' 'adverse' 'advisory' 'aerial' 'affect' 'affected'
 'africa' 'after']
```

Figure 14 – First 10 rows and columns showing TF-IDF scores for lemmatized headline data

	abating	ability	able	abnormally	about	above	academic	access	accom	\
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figure 15 – First 30 vocabulary features after stemming and vectorization on headline data

```
['abat' 'abil' 'abl' 'abnorm' 'about' 'abov' 'academ' 'access' 'accom'
 'accord' 'accordi' 'account' 'accus' 'acr' 'acropoli' 'across' 'action'
 'activist' 'adam' 'adapt' 'address' 'adriana' 'advent' 'advers'
 'advisori' 'aerial' 'affect' 'africa' 'after' 'afternoon']
```

Figure 16 – First 10 rows and columns showing TF-IDF scores for stemmed headline data

	abat	abil	abl	abnorm	about	abov	academ	access	accom	accord	...	\
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	

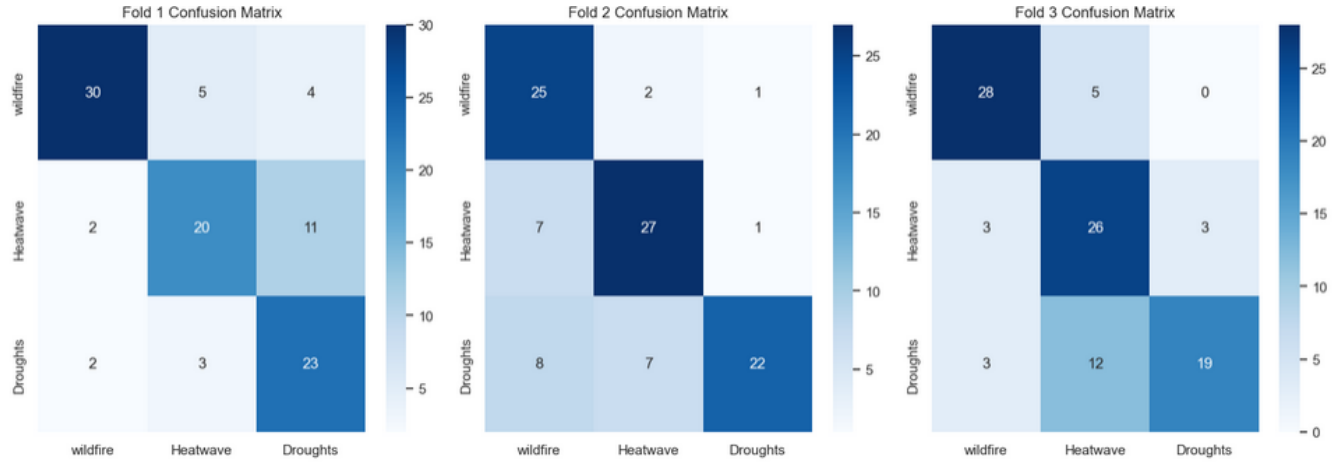
Model 1: Multinomial Naive Bayes Results (3 Cross Validation):

Prior to initializing the Multinomial Naive Bayes (MNB) model with the vectorized headline data, a three-fold data splitting approach was adopted. This strategy involved segmenting the dataset into three distinct subsets, each comprising a training set and a corresponding test set. Within this framework, the relevant columns 'df_processed' and 'LABEL' were isolated and utilized as the feature variable X, while the target variable 'LABEL' was designated as y.

Upon applying the MNB model to this prepared dataset, an aggregate accuracy of 73% was achieved. Figure 17 visually presents the outcomes of the confusion matrix analysis for each of the individual folds. In Fold 1, the model demonstrated its highest predictive prowess with regard to wildfire articles, followed by heatwaves and droughts in second and third positions, respectively. In Fold 2, heatwave predictions led the performance, succeeded by wildfire and drought predictions. For Fold 3, wildfire articles again secured the top position, with heatwaves in second and droughts in third.

The overarching trend revealed that the MNB model excelled in its ability to accurately predict articles categorized under the wildfire topic. However, when faced with articles pertaining to droughts, the model encountered greater difficulty. This discernment highlights the model's varying proficiency across different weather-related topics and underscores the complexity in predicting articles belonging to the droughts category.

Figure 17 – Confusion Matrix heatmap for MNB 3CV



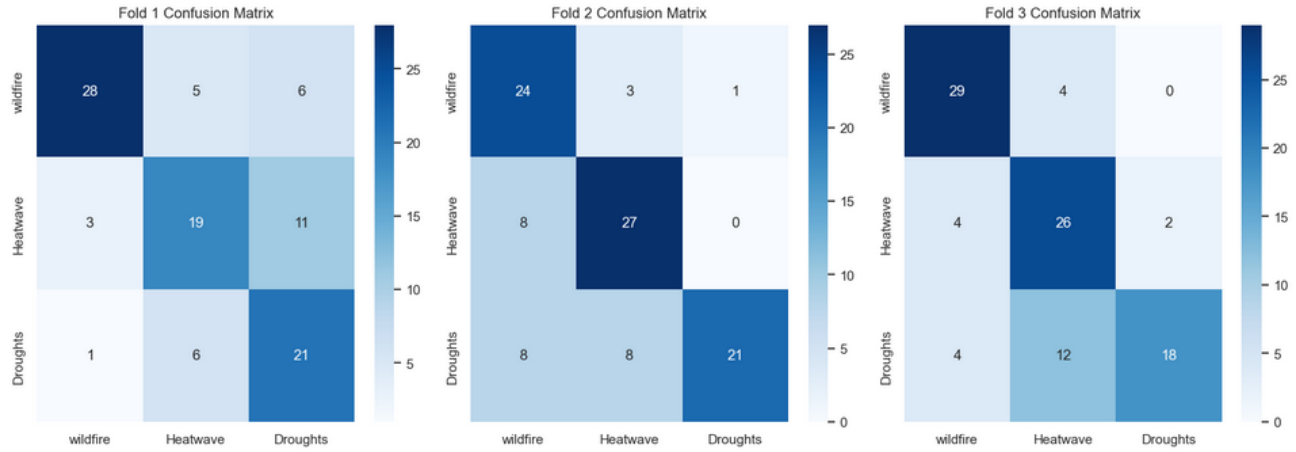

```
Cross-validation scores: [0.78      0.72      0.6969697]
Mean accuracy: 0.7323232323232324
```

Model 2: Naive Bayes + Lemmatization

In the second model iteration, lemmatization was integrated into the text preprocessing pipeline. Lemmatization served to standardize words by reducing inflections and obtaining their base forms. Similarly to the prior models, the second model's evaluation comprised 3-fold cross-validation. Despite the incorporation of lemmatization, the outcome remained consistent with the initial model run.

Across all three cross-validation folds, the same classification patterns persisted. Notably, wildfire-related articles continued to exhibit the highest classification accuracy, while drought-related articles maintained their position as the least accurately predicted category. This consistency underscores the robustness of the classification system and reaffirms the distinctiveness of the different weather-related topics in the dataset.

Figure 18 – Confusion Matrix heatmap for MNB 3CV + Lemmatization



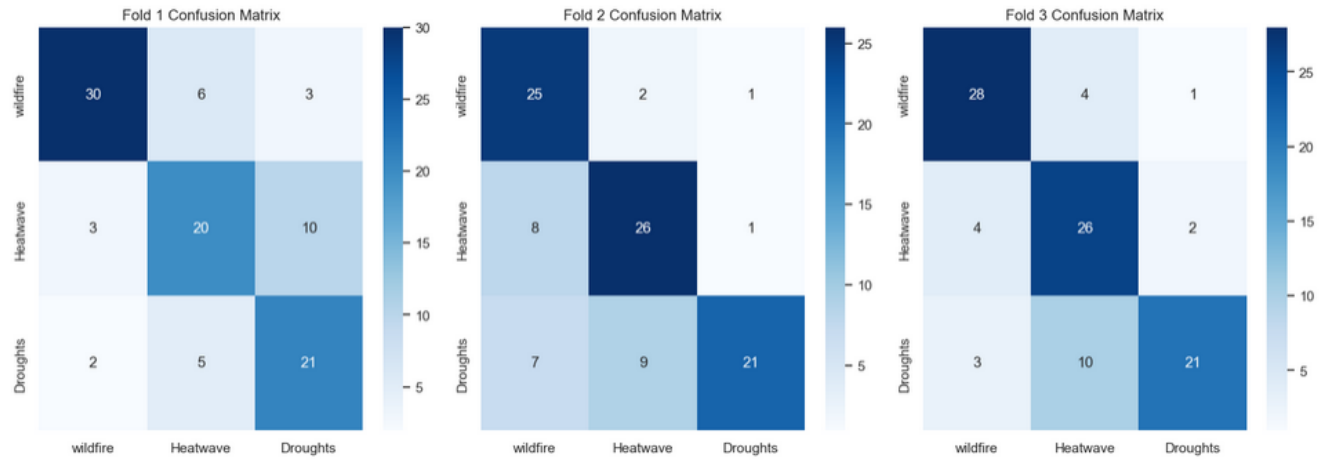

```
Cross-validation scores: [0.8        0.73       0.67676768]
Mean accuracy: 0.7355892255892256
```

Model 3: Naive Bayes + Stemming

In the third model iteration, an alternative text processing technique called stemming was employed instead of lemmatization. Stemming involves truncating prefixes and suffixes from words, aiming to further simplify them. Like the initial two models, the third model's performance was assessed using 3-fold cross-validation. Notably, a slight improvement in overall accuracy emerged, with the accuracy score increasing from 73.56% to 75.23%.

Despite the accuracy enhancement, the relative rankings among the classes remained consistent. Wildfire-related articles continued to exhibit the highest prediction accuracy, followed by heatwave-related articles. Drought-related articles maintained their position as the third most accurately predicted class. This observation highlights the robustness of the class distribution and the model's ability to distinguish between these distinct weather-related topics.

Figure 19 – Confusion Matrix heatmap for MNB 3CV + Stemming



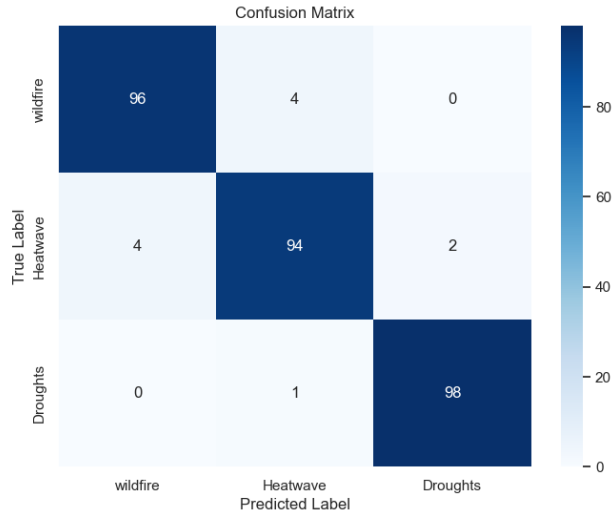

```
Cross-validation scores: [0.81       0.75       0.6969697]
Mean accuracy: 0.7523232323232323
```

Model 4: Random Forest + Lemmatization

The fourth model introduced a novel approach by employing the Random Forest algorithm in conjunction with TF-IDF vectorization and lemmatized text data. The outcomes of this combined methodology are indeed noteworthy. Notably, the overall accuracy, which previously stood at 75.23% in the third model, experienced a remarkable surge, soaring to an impressive 96.32%.

This substantial enhancement in accuracy is consistent across all classes, with the performance of the droughts category slightly surpassing that of the wildfire and heatwave categories. This achievement is a testament to the potent capabilities of the Random Forest algorithm when harnessed with TF-IDF vectorization and lemmatized text data. The algorithm's ability to address the intricate challenges inherent in severe weather event classification is clearly demonstrated through these highly accurate results. This significant advancement holds the potential to revolutionize the field by offering a robust solution for discerning and categorizing severe weather news articles with precision.

Figure 20 – Confusion Matrix heatmap for Random Forest + Stemming



Accuracy: 0.9632107023411371

Conclusions

In the realm of severe weather news classification, the analysis has unveiled significant findings that can impact a wide range of individuals and organizations. By delving into the complexities of identifying and categorizing news articles related to severe weather events, this study aimed to uncover meaningful insights that transcend technical jargon. The objective was to offer accessible information that can guide decision-making, influence policies, and enhance preparedness for a diverse audience.

The study's outcomes extend beyond technical domains, providing practical implications for people from various backgrounds. From concerned citizens seeking real-time updates on weather conditions to emergency response teams making critical choices, the analysis illuminates trends and patterns within severe weather news. By recognizing prevalent themes like storm wildfire or heatwave reports and identifying consistencies in news coverage, this research contributes to informed decision-making at both personal and professional levels.

Moreover, the study underscores the potential for expanding topic classification to encompass broader domains. The techniques employed in this analysis, refined and adapted, can serve as a foundation for categorizing news articles across various subjects beyond severe weather. This approach could aid in enhancing the accessibility of news, allowing users to quickly find relevant information amidst the overwhelming volume of data. By presenting these conclusions in an easily understandable manner, the research seeks to encourage discussions and collaborations that push the boundaries of topic classification, ultimately enabling a wider audience to harness the benefits of structured news categorization.