US Religious and Political Landscape and the Effect on UFO Sighting Reports Sentiment and Counts

Stephanie Buchanan, Thad Hoskins

November 17, 2021

1 Executive Summary

Introduction On December 12, 2020, Congress passed a $2.3 trillion Omnibus Appropriations and Coronavirus Relief Package that included the requirement for the Pentagon and other agencies to provide a report of all known information about UAPs (unidentified aerial phenomena) previously known as UFOs (unidentified flying objects). A large volume of UFO sightings have been attributed to pranks and misidentified Starlink satellites or rare military aircraft. This analysis described herein explores whether the religious or political affiliations within the U.S. plays a role in the sentiment or count of UFO sighting reports reported to the National UFO Reporting Center (NUFORC).

The first analysis is to generally understand the data. For this purpose, we employed NLP to evaluate the sentiment and objectivity of the sightings, as well as less complex analysis of words used and other descriptors provided for each sighting. Next, we mapped the data using the geographical features provided to gain an understanding of visual patterns across the sightings to specifically address whether sightings skew toward geographical regions.

Research Questions The goal of this study is to answer the following research questions using data mining, machine learning and NLP techniques:

- Is there a relationship between the religious and political affiliation of a state and UFO reports?
  - regarding the sentiment of sighting reports?
  - regarding the number of sightings reported?
- Which sentiment analyzer performs best with the UFO sighting reports?
- When and where are you most likely to view a UFO?
- Are there any interesting trends observed in religious or political affiliation between the years of 2007 and 2014?

Actionable Insights Three sentiment analyzers, VaDER, Nltk, and TextBlob, were compared and the VaDER sentiment analyzer was found to classify a wider range of sentiment than the

other two. Thus, the VaDER sentiment analyzer is recommended in the case of analyzing the sentiment of UFO sighting reports.

Regardless of opinions regarding the validity of UFO sightings, simplistic views of religion and political views does not inform one regarding evaluation of sentiment and numbers of sightings.

The probability of a sighting report increases on the 4th of July and the 21st of December. Most of the UFO sightings were reported to occur between the hours of 4 to 5 AM, so this time is also recommended. Many reports originate in the Southwest region of the U.S.

The religion distributions of all states stayed similar between the years of 2007 and 2014 except in Massachusetts and Rhode Island. The number one religion in 2007 in both of these states was Mormonism, but in 2014 switched to Judaism. This analysis did not explore the reason behind the switch, but this observation may be researched to determine root cause.

The political data showed the trend of increased numbers in the Democrat and Republican parties between the years of 2007 and 2014. Very few people responded 'Don't know' or 'No preference' in 2014, compared to higher numbers in 2007. This increase in the number within the 'large' parties may be of interest to another research group, but was not explored further in this analysis.

2 UFO Sighting Reports Dataset

The data was scraped from NUFORC's website. The website was setup as a way for people to self-report UFO sightings around the world. There have been a large volume of sightings over the past year. The site attributes a lot of these sightings to the Starlink satellites, and cautions researchers to be discriminating with the information being presented in the reports.

2.1 Scraping the Data

For our analysis, we wanted to use Natural Language Processing, or NLP, on the data sets we evaluated. The text of the sightings was truncated, as was the "list view" in the website. To get the full summary of the sighting, one needs to use the detailed summary page. Since this was not available, it was decided to go to the source.

A quick look at the list of sightings by date (NUFORC Index by Month) confirms a large number of records. In this case, they are broken down by month and year. The vast majority of the sightings are from the last 20 years, but that affects the scrape little since we want them all.

In total, the scraping of the data was a five day process but all of the data provided by NUFORC was obtained. The details of the scraping may be provided in another report but was not as pertinent to this project.

2.2 Data Cleaning

The columns in the original dataset before cleaning were "Date_Time", "City", "State", "Shape", "Duration", "Summary", "Posted", "Detail_Link", and "Detail_Summary". The "Detail_Summary" column was populated from the "Detail_Link" if available, and if not, was populated from the "Summary" column.

City, County, State, Country After exploring the data, many city fields contained information that was additional to the city name. Some contained parenthetical descriptions that included Country, comments, hotel names, etc. Likewise for the state which contain state, regional, country, oceans, etc.

Knowing we would limit our scope to the United States, anything that could not be identified positively as being in the United States was discarded. It is likely there are US sightings in that data, but with 65,000 remaining records the cost to benefit ratio pointed to letting them go.

The location data was cleaned down to a city and state, then using available ARCGIS library, Latititude and Longtitude was added to the dataset. Further, the county where the city is located was added. As we will explore later, we are now able to aggregate the data to the granular city, county, and state levels.

Date_Time We encountered an issue when converting the "YY" format of the year to a "YYYY" format. All years in the 1900's up to 1967 were correctly formatted with "19" as the first two digits. But from 1968 on, the years were formatted with "20" as the first two digits. We applied a function to this column to subtract 100 from any year greater than 2021.

Finally, all times were converted to UTC. This was achieved by first identifying the timezone using the latitude and longitude and the TimezoneFinder() function. The timezone was then used to convert into UTC using the .astimezone(pytz.utc) function.

Detail_Summary Because our primary application is NLP, cleaning the text is important. Standard text cleaning functions were applied including removing stop words, expanding contractions, and lemmatization. The NLTK library was utilized for the available built-in functions.

The order of the operations is important to ensure consistent cleaning of text. The data was first made lower case to standardize on lower. Next, the text was tokenized to separate individual words for processing. Directly after tokenizing, the string was joined back together with stop words removed. Regular expressions were used to remove newline characters, punctuation and special symbols. Hyphens were replaced with white spaces, then stripped of any white spaces from the beginning and end of each string.

Next, contractions were expanded using the built-in 'contractions' NLTK library.

Lastly, we Lemmatized the words, converting the words to their root word, e.g., changing run, ran, running to run.

The cleaned text were saved to an additional column with the original detailed summary carried

further for use in the updated VADER sentiment analyzer.

Shape Shape is an attribute we read directly from the NUFORC data. We standardized similar names, such as "triangular" to "triangle," "rectangular" to "rectangle," etc.

2.3 Visualizing Sightings Geographic Data

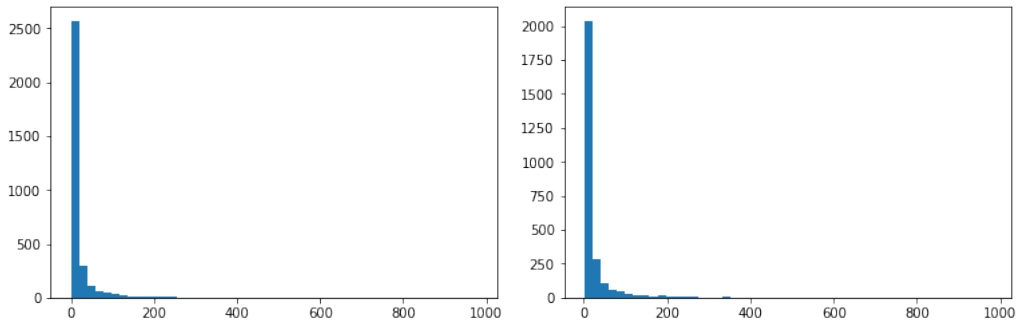
Other than the sightings details the next major component of the sightings is the location. As mentioned above, the city and state were cleaned such that the remaining records were clearly noted. For the purpose of mapping sighting counts, the county for each city was added and the number of sightings were aggregated to that level. This creates an easier to interpret map, not too granular (city), not too summary (state).

Three values are in the dataset for the number of sightings:

- raw number of sightings per county
- the number of sightings per one thousand residents
- "smoothed" (the log of the value) number of sightings per thousand residents

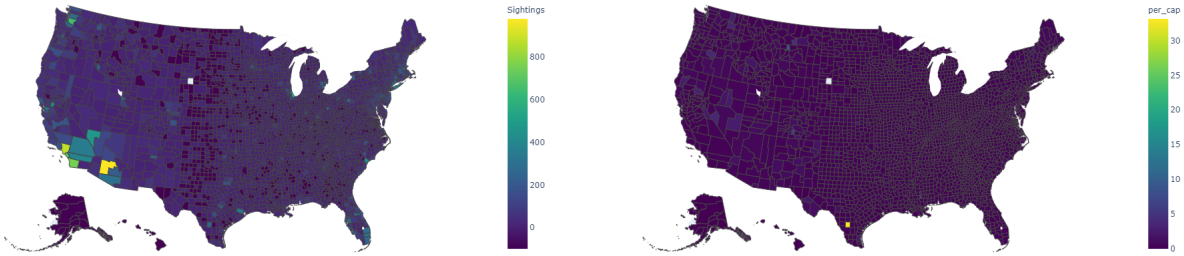
As the histogram shows below, the number of sightings per county is very low with a mean of 18.76, median of 4. For context in the figure, there are a total of 3,220 counties in the dataset with 553 having no sightings.

Figure 1: Sightings by County Histogram (left: including counties with 0 sightings; right: at least 1 sighting)



The "sightings per thousand residents" paints a similar picture of sightings with a mean of 0.25346, median of 0.15754, and a maximum value of 33.11.

Figure 2: County Raw Number of Sightings of sightings per thousand residents (-100 represents no sightings; number scaled for color contrast)



La Salle, Texas With a population of 7520, La Salle County in Texas has the highest number of sightings per thousand residents at 33.111702, which is 249 sightings. Walthall County, Mississippi, with a population of 14,286, comes in second with 124 total sightings and 8.6798 per thousand residents.

La Salle can be clearly seen in the above "sightings per thousand residents" figure in the bright yellow square in southern Texas.

Maricopa County, Arizona and Los Angeles County, California Maricopa County has the highest number of sightings in total in this dataset with 979. Other datasets report that Maricopa County has the most number of sightings in the United States, but this title is contested. Regardless, Maricopa County is highlighted above in the map of raw numbers of sightings.

This may be accounted for with its proximity to Area 51 and its proximity to heavily populated areas, such as Los Angeles, which competes with Maricopa County for the most number of sightings in other datasets. In this dataset, Los Angeles County places second in the number of sightings with 877 to Maricopa's 979. The counties around Los Angeles County, including Los Angeles account for 3,216 sightings.

The Los Angeles County area mentioned above along with Maricopa County and surrounding counties account for 4,709 sightings or 7.795% of sightings in this dataset.

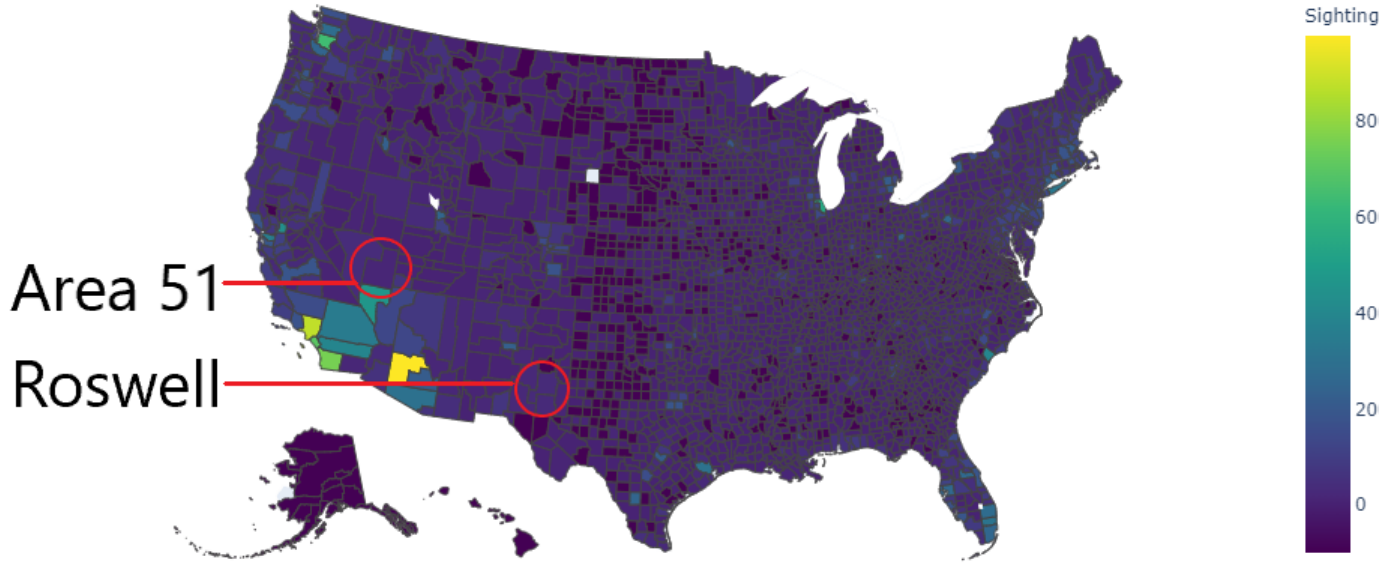
Other Notable Sites Area 51 (Lincoln County, Nevada) and Roswell (Chaves County, New Mexico) are notable locations to track. The figure below highlights both. Nearby Clark County accounts for 466 sightings.

The county in which Area 51 sits has recorded 13 UFO sightings. Surrounding counties, including Clark County, Nevada record approximately one hundred sightings.

Similarly, the county housing the International UFO Museum in Roswell, New Mexico claims 29 sightings, while Lincoln County, New Mexico where the Roswell incident of the 1947 occurred, accounts for 23 sightings.

To complete the Top 5, We move up the west coast to King County, Washington which has 643 sightings.

Figure 3: Notable UFO Sites

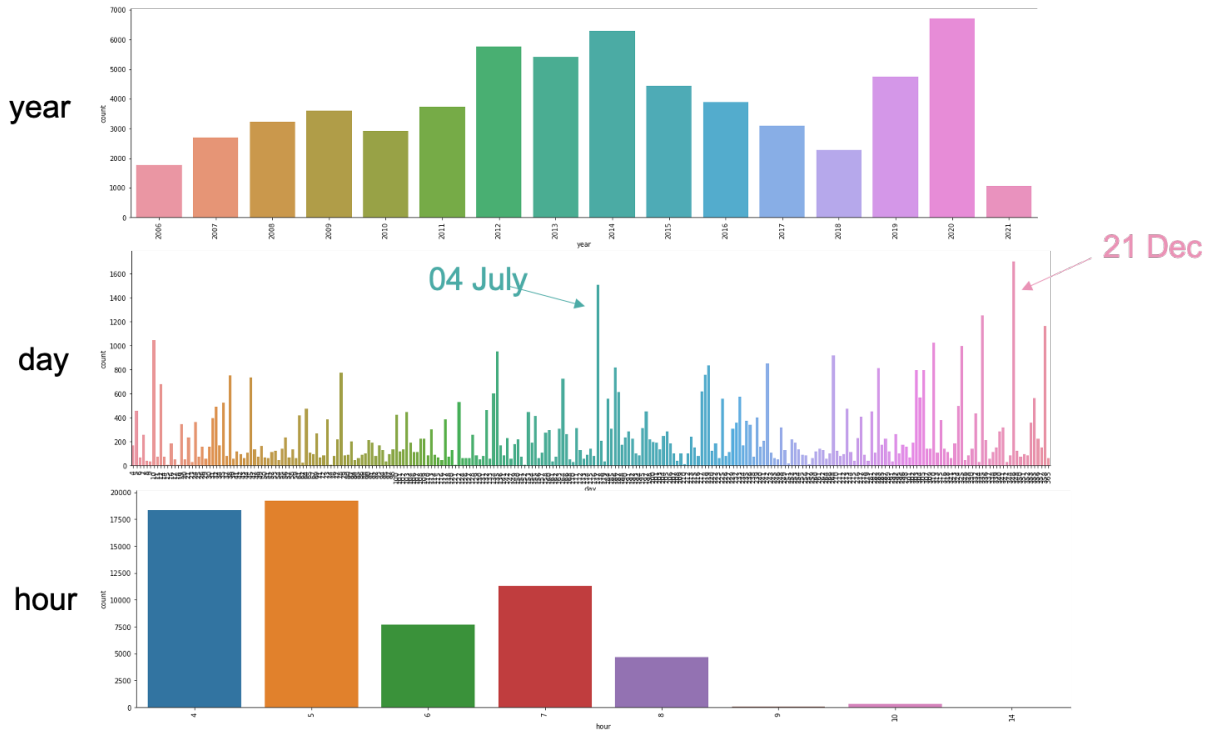


3 Count trends

To understand the trend in counts of UFO sightings between the years of 2006 and 2020, count plots were generated by year, day and hour in UTC time. These plots are shown in Figure 4 and show:

- 2014 and 2020 had the most reported sightings,
- the 4th of July and 21 December are popular days for sightings, and
- most sightings within this year range occurred between the hours of 4 to 5 AM.

Figure 4: UFO Sighting Count Trends by Year, Day and Hour



4 Pew Research Center Datasets

The religious and political affiliation datasets were downloaded from the Pew Research Center available data sets. The two datasets that were used in this analysis were the 1) 2007 and 2)2014 U.S. Religious Landscape Surveys.

4.1 2007 U.S. Religious Landscape Survey

In the summer of 2007, the Pew Forum on Religion and Public Life conducted the main part of the Religious Landscape Survey with a representative sample of 35,556 adults living in the continental United States. In the spring of 2008, the Pew Forum conducted a supplemental survey with a sample of 200 adults living in Alaska and 201 adults living in Hawaii. This analysis focused only on the contiguous 48 states to keep the timing consistent since analyzing the differences across the 2007 and 2014 surveys was of interest.

The 2007 survey was conducted over the phone by Princeton Survey Research Associates International (PSRAI) in English and in Spanish. Standard list-assisted random digit dialing (RDD) was used in selecting the survey participants. An additional 547 were surveyed that were identified as being Hindu, Buddhist or Orthodox Christian. This helped boost the sample size for these religions that are considered 'low-incidence'.

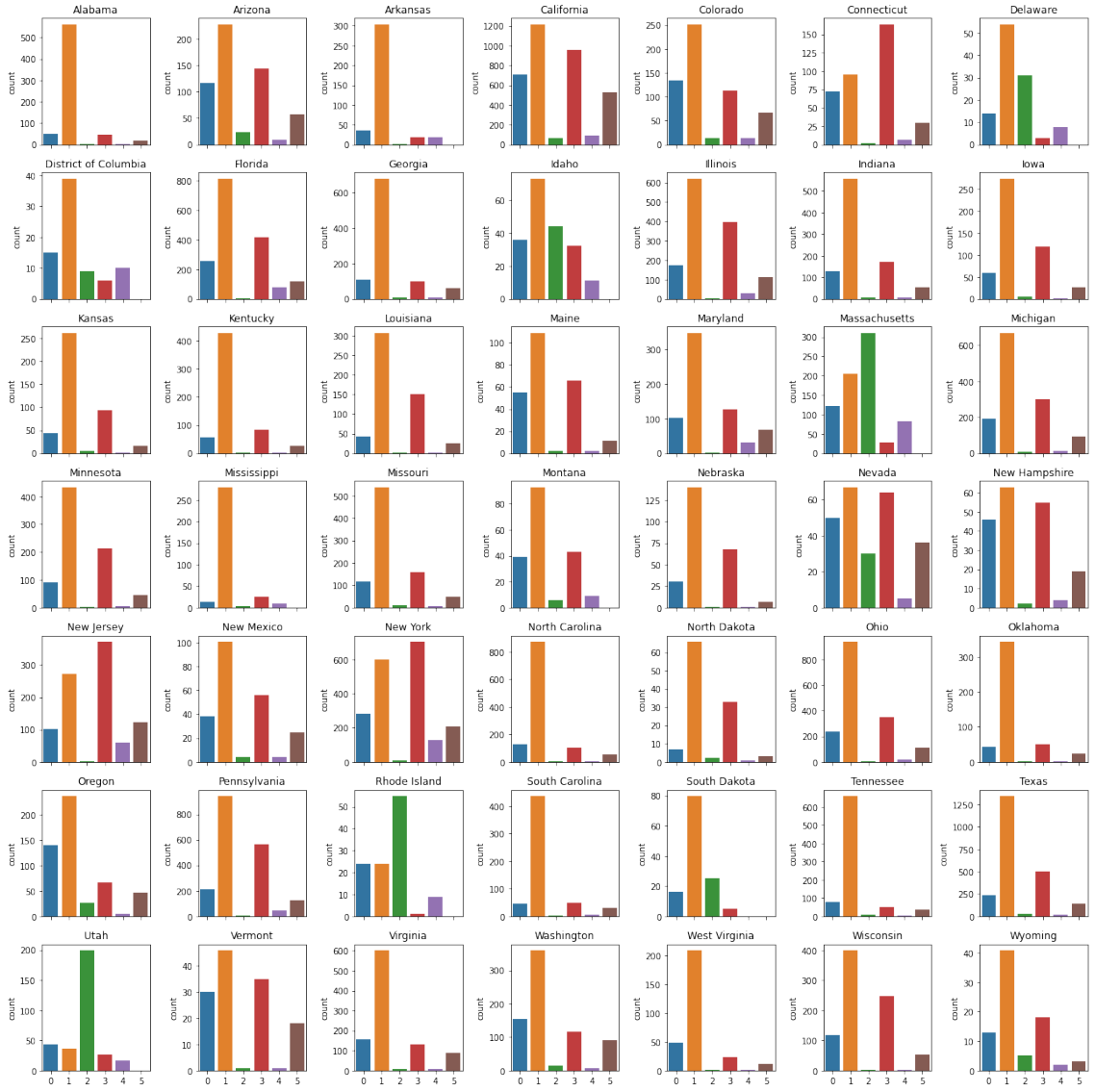
There were 65 options for religion in the 2007 and 2014 surveys. After some visualizations, it was seen that the most frequent responses were 'Protestant', 'Roman Catholic', and 'Mormon

(Church of Jesus Christ of Latter-day Saints/LDS)'. To reduce the complexity the following labels were defined and utilized:

- if 'Nothing in particular', 'Atheist (do not believe in God)', or 'Agnostic (not sure if there is a God)' or x=='Nihilist (VOL)': 0
- 'Protestant': 1
- 'Mormon (Church of Jesus Christ of Latter-day Saints/LDS)': 2
- 'Roman Catholic': 3
- Jewish (Judaism)': 4
- 'Other': 5

Figure 5 shows the distributions of these particular religions across each of the 48 contiguous states along with Washington D.C in 2007. In every state except Connecticut, Massachusetts, New Jersey, New York, Rhode Island, and Utah, Protestant was the most common religion. In Massachusetts, Rhode Island and Utah, Mormon was the most common religion.

Figure 5: Religion Distribution in 2007 Across the Contiguous U.S.

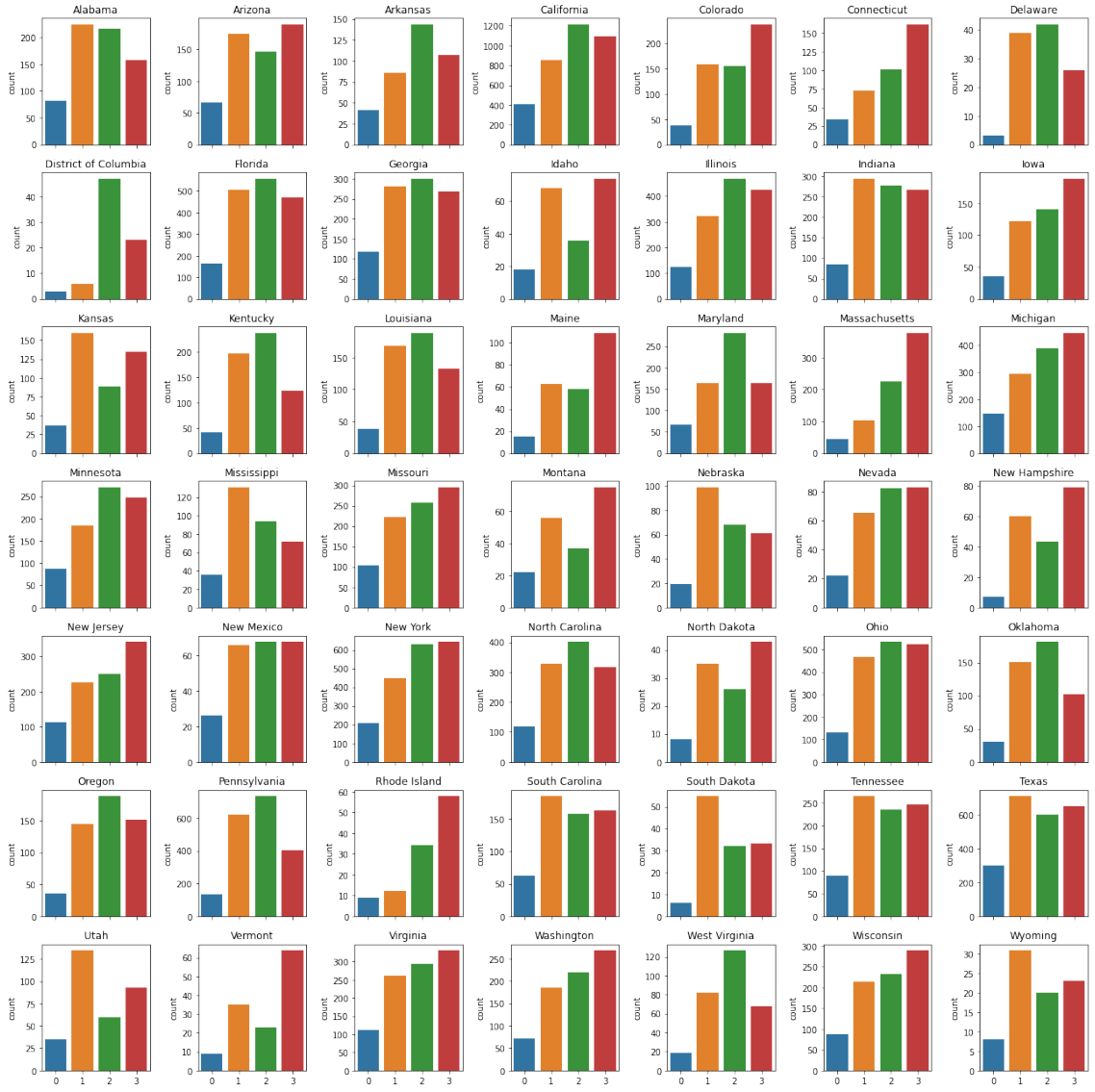


Along with the survey question asking the participant to identify their religious affiliation, the political affiliation was also asked. Since this was of interest, we pulled this data to use in our models. The distribution of political affiliation is shown in Figure 6. The following labels were used:

- 'Don't know/Refused' or 'No Preference': 0
- 'Republican': 1
- 'Democrat': 2
- 'Independent':3

The 'Don't know/Refused' category was fairly low in all states except in Michigan, Minnesota, South Carolina, Virginia and Washington.

Figure 6: Political Affiliation in 2007 Across the Contiguous U.S.

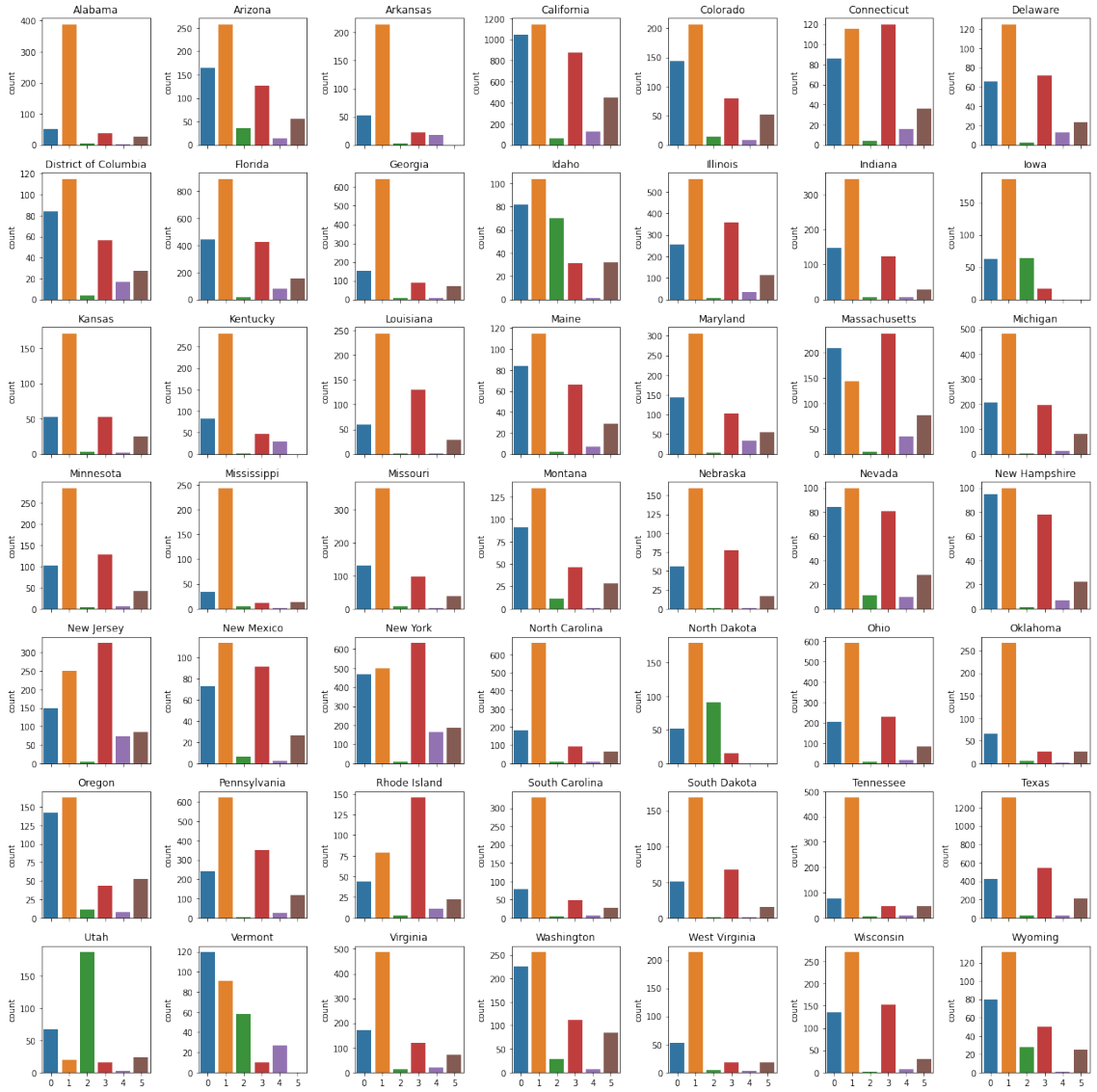


4.2 2014 U.S. Religious Landscape Survey

The 2014 U.S. Religious Landscape Survey was conducted in the summer of 2014 among a sample of 35,071 U.S. adults. The surveys were conducted over the phone, and approximately 60% of the interviews by cellphone (n=21,160) and 40% were by landline (n=13,911). A minimum of 300 interviews were conducted in every state and the District of Columbia, and the survey is estimated to cover 97% of the U.S. adult population.

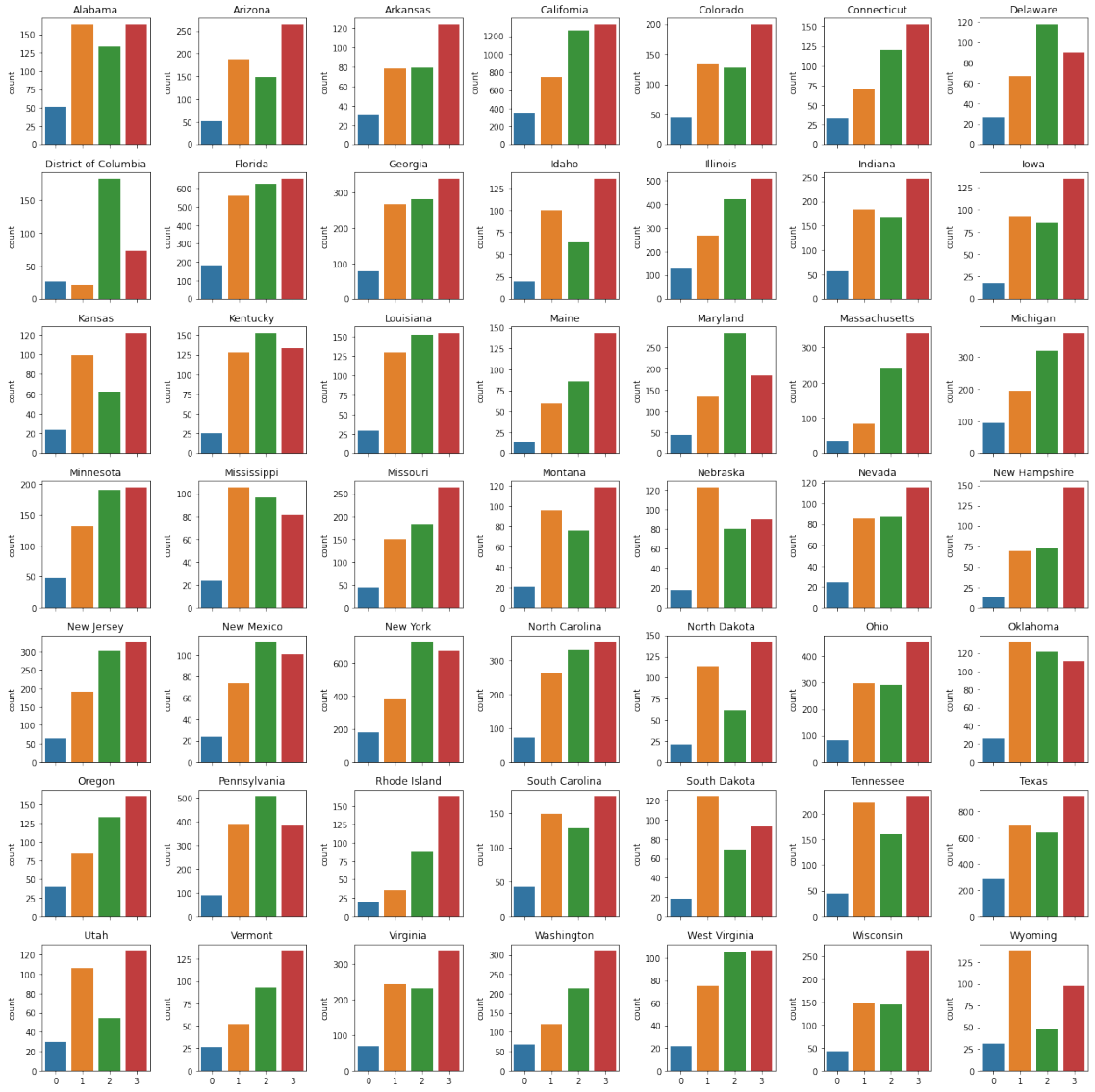
The same labeling as described above was used for the 2014 dataset. Figure 7 shows the distributions of these religions across each of the 48 contiguous states along with Washington D.C in 2014. In every state except Connecticut, Massachusetts, New Jersey, New York, Rhode Island, Utah, and Vermont, Protestant was the most common religion. Unlike in 2007, only in Utah was Mormon was the most common religion.

Figure 7: Religion Distribution in 2014 Across the Contiguous U.S.



The distribution of political affiliation in 2014 is shown in Figure 8. Unlike in 2014, the 'Don't know/Refused' category was fairly low in all states reflecting a political polarization over the years between 2007 and 2014 with more people identifying with a political party.

Figure 8: Political Affiliation in 2014 Across the Contiguous U.S.



Feature Transformation In both the 2007 and 2014 datasets, the labeled religion and party features were converted into a frequency of each religion and party by state. This was done to normalize the values into the range of 0 to 1. The frequencies of each religion and party was then exploded out into separate columns by state. These were the features used in the machine learning models described below.

5 Technique Summary

5.1 Natural Language Processing

NLP is described as a branch of artificial intelligence with roots in computational linguistics[1]. These techniques help computers understand, analyze and manipulate human languages. For

[1]source: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

example, NLP makes it possible for a computer to read text, interpret it, determine the most common phrases, measure sentiment, and measure objectivity of the text.

The main platform used in the analysis of the UFO sightings dataset was the Natural Language Toolkit[2] This toolkit provides a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

An 'n-gram' is defined as a sequence of n words in some text. The n-grams that interest us are common word groupings in the UFO sighting summaries.

5.1.1 Text Exploratory Analysis

Summary Statistics The UFO sighting reports were explored to identify common characteristics and any anomalies present in the dataset. The count of sentences, words, characters, and stopwords were calculated, and then the histograms of each of these features were plotted.[3]

Figure 9 shows the statistical summary of each of the distributions. As seen in the data, the average number of sentences in the sighting reports is 1.29, average number of words in each report is 14.03, average number of characters is 67.35 and average number of stopwords is 4.76.

Figure 9: Statistical Summaries of Text Features

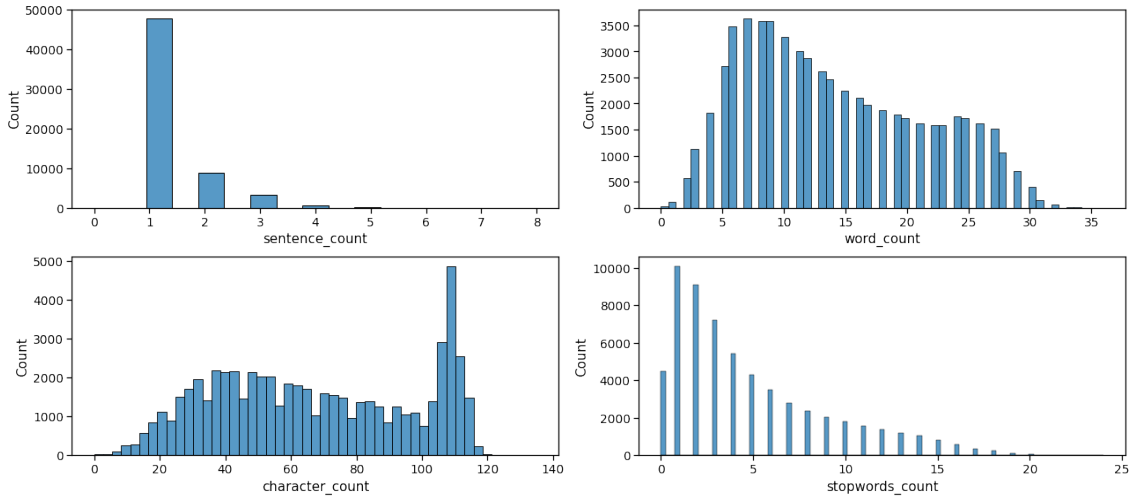
	sentence_count	word_count	character_count	stopwords_count
count	60412.000000	60412.000000	60412.000000	60412.000000
mean	1.289032	14.026667	67.345610	4.759237
std	0.635039	7.332038	30.249404	4.191235
min	0.000000	0.000000	0.000000	0.000000
25%	1.000000	8.000000	42.000000	2.000000
50%	1.000000	13.000000	64.000000	3.000000
75%	1.000000	20.000000	97.000000	7.000000
max	8.000000	36.000000	135.000000	24.000000

Figure 10 shows the distributions of each of these text features. The sentence and stopwords counts show left-skewed distributions in the data with 75% of the reports having 1 sentence and 7 or fewer stopwords. The word count shows a fairly normal distribution. The character count seems to show two normal distributions with one centered around 50 characters, and the other around 110 characters.

[2]Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

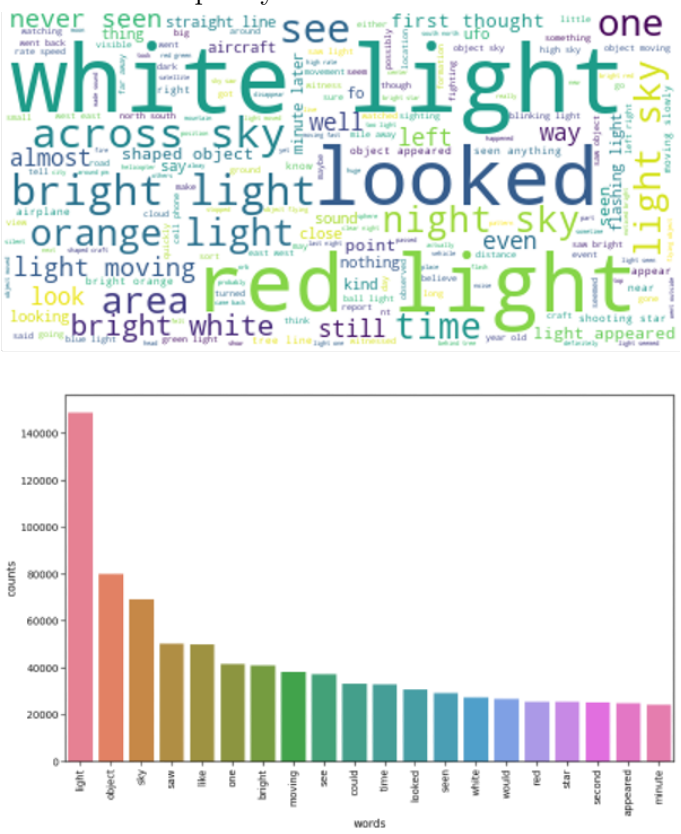
[3]https://medium.com/swlh/text-summarization-guide-exploratory-data-analysis-on-text-data-4e22ce2dd6ad

Figure 10: Distribution of Text Features



Word Frequencies The most frequent words present in the sighting reports were explored using wordclouds and the CountVectorizer() found in the sklearn.feature_extraction.text library. The WordCloud shown in Figure 6 show the most common words and n-grams in the UFO sighting reports are: looked, white light and red light. The CountVectorizer converts the collection of UFO sighting reports to a matrix of token counts. The counts of the 20 most common words are shown in the count vs word plot in Figure 11. The most common word in the reports by far is 'light'.

Figure 11: Word Frequency via WordCloud and CountVectorizer



Topic Modeling Topic modeling was explored in this analysis to identity topics that are discussed in the UFO sighting reports. Topic modeling is an unsupervised machine learning technique used in natural language processing. A 'topic model' is a type of statistical model used to cluster groups of semantically similar words for topic representation. The model detects patterns such as word frequency, and distance between words.

Latent Dirichlet Allocation (LDA)[4] is a topic modeling technique that converts the set of sighting reports to a set of topics. LDA is a three-level hierarchical Bayesian model, where each word in the corpus is modeled as a finite mixture over an underlying set of topics. The goal of LDA is to determine in which topic cluster a document belongs using the words present in the document. The scikit-learn and pyLDAvis libraries were used for the topic modeling analysis.

The scikit-learn library provides a function to calculate LDA and returns a list with the topic and the tokens that include. In our analysis, we define 10 topics to identify and then print out the top 12 words in each topic from our corpus. The topics and top words are shown in Figure 12.

[4]S.S., R. and Dr.P., P. (2018). Topic Categorization on Social Network Using Latent Dirichlet Allocation. Bonfring International Journal of Software Engineering and Soft Computing, 8(2), pp.16–20

Figure 12: Ten Topics and the Top 12 Words Associated with Each Topic.

```
Topics found via LDA:

Topic #1:
craft light triangle one shape three two triangular shaped covering covered white

Topic #2:
like saw see could looked back time seen would thing went know

Topic #3:
light car driving saw ufo road side tree home bright right red

Topic #4:
time approximately would mile degree area observed lake appeared sighting year point

Topic #5:
ball fireball fire firework falling sky trail july like tail method flame

Topic #6:
light sky bright star red moving white like saw one looked night

Topic #7:
light orange one sky moving north south east west two formation line

Topic #8:
flying aircraft plane jet sound low helicopter airport air speed could flight

Topic #9:
object sky cloud appeared moving shape second shaped like moved white large

Topic #10:
video picture phone camera took photo object take get fo cell one
```

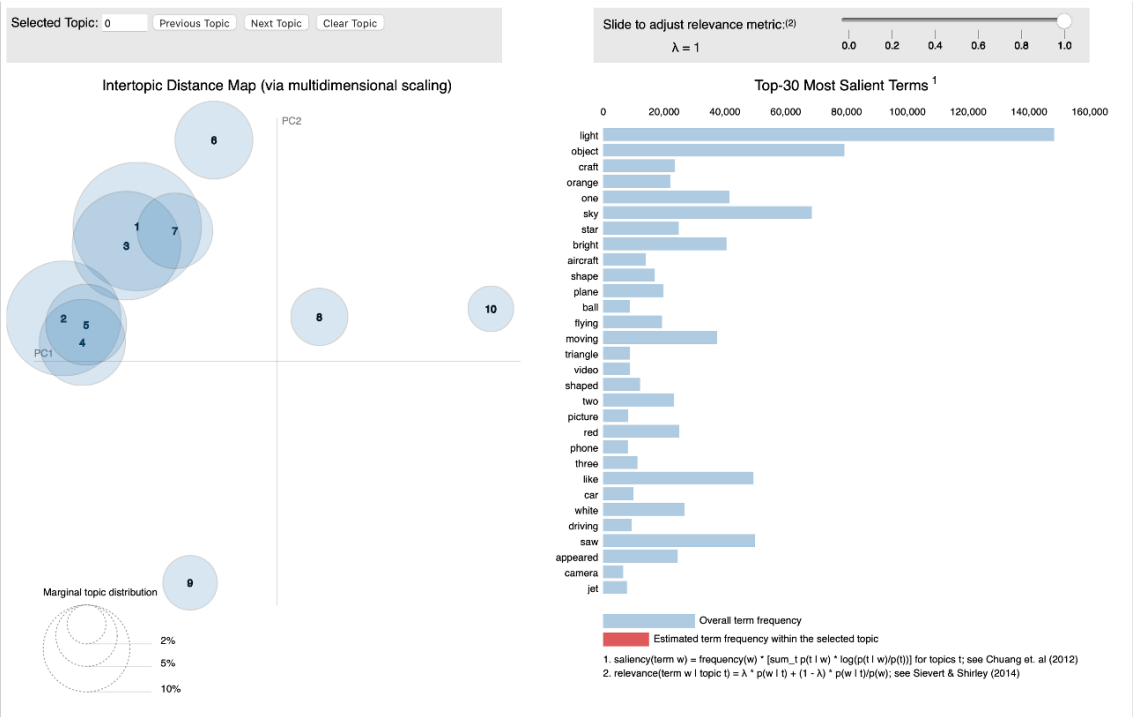
To visualize the topics, the pyLDAvis library was used to generate an interactive figure where the size of the bubble represents the importance in the text contained in the cluster. The distance between each circle represents the similarities of the topics. The parameter λ is a relevance metric that distinguishes words which are exclusive to the topic, closer to 0, rather than words with high probability of being included in the selected topic which would be closer to 1. Exploring different values of λ for each topic can assist in assigning a name to each topic.

Figure 13 shows the Intertopic Distance Map via multidimensional scaling and the top-30 terms across the entire corpus. There is an interactive feature that allows the user to select a topic and view the most relevant terms in that particular topic while adjusting the relevance metric.

As seen in the Intertopic Distanse map, there are 3 main clusters, 1 containing topics 1 through 7, and then separately, topic 8, topic 9 and topic 10. The 3 most common words in each of these separate topics are:

- Topic 8:'video', 'picture', 'phone'
- Topic 9:'triangle', triangular', 'saucer'
- Topic 10: 'ball', 'fireball', 'fire'

Figure 13: Topic Modeling via LDA Visualization.



Sentiment analysis Sentiment analysis can help determine the ratio of positive to negative engagements in a group of words, in this case, detailed summaries of UFO sightings. Algorithms are used to classify text into positive and negative categories. For this project, three pre-trained sentiment analyzers were compared: NLTK's sentiment analyzer called VADER (Valence Aware Dictionary and sEntiment Reasoner) was used on two features: a) the raw detailed summary and b) the cleaned detailed summary, and c) TextBlob sentiment analyzer on the cleaned detailed summary.

VADER is best suited for short sentences with some slang and abbreviations like the text commonly found in social media. For this reason, the summaries were split into sentences and analyzed separately. The output from VADER is a dictionary of different scores: negative, neutral, positive, and compound scores. The negative, positive and neutral scores all add up to 1 and can't be negative. The compound score is similar to an average, but calculated differently and can range between -1 and 1. Recent improvements to the VADER analyzer allows for uncleaned data to be input and analyzed. This version was used to understand the differences between capitalized letters and lowercase.

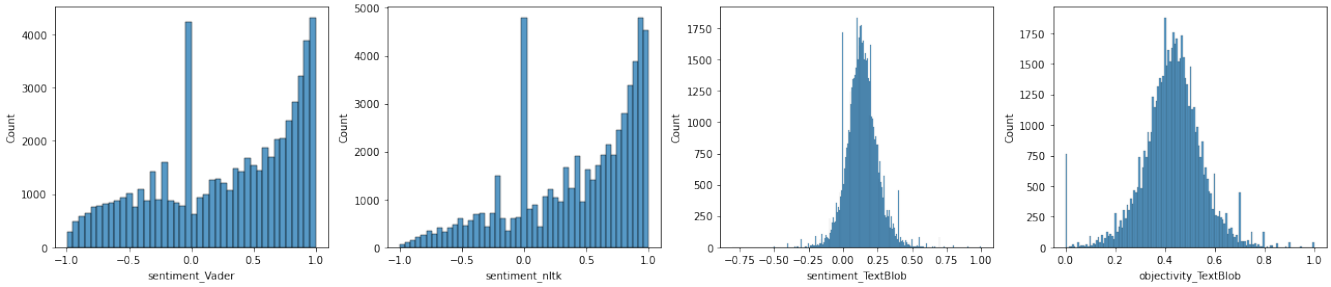
The TextBlob library provides a function to determine the sentiment of a sentence along with the objectivity of a sentence. The output of the TexBlob sentiment function is a tuple of two values: polarity and subjectivity. Like the compound score from NLTK's sentiment analyzer, polarity is a float value within the range [-1.0 to 1.0] where 0 indicates neutral, +1 indicates a very positive sentiment and -1 represents a very negative sentiment.

Subjectivity is a float value within the range [0.0 to 1.0] where 0.0 is very objective and 1.0 is very subjective. A subjective sentence will express personal feelings, views, beliefs, opinions,

allegations, desires, beliefs, suspicions, and speculations. In contrast, an objective sentence is factual.[5]

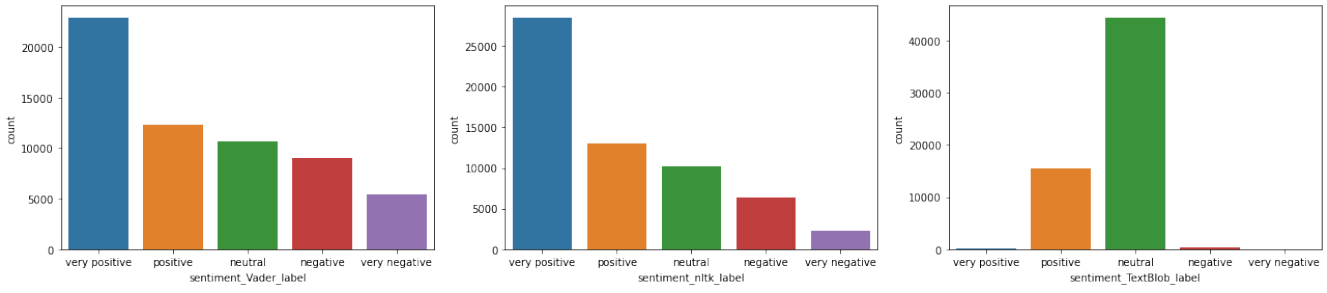
Figure 14 shows the distribution of the raw sentiment values using each version of sentiment analyzers. Objectivity is also included for comparison. The sentiment analyzer using the 1) raw Detailed Summary, 'sentiment _Vader', and 2) cleaned detailed summary, 'sentiment_nltk', show very similar distributions with many reports at 0.0, and a skewed-right distribution. The TextBlob sentiment, sentiment_TextBlob, and objectivity, 'objectivity_TextBlob', values show normal distributions with very few outliers.

Figure 14: Continuous Values Distribution



To visualize the differences between the sentiment analyzers, the values were bucketized into 5 buckets: 'very positive', 'positive', 'neutral', 'negative', and 'very negative'. These distributions are shown in Figure 15. As seen in the data, subtle different exist between the VADER sentiment analyzer when using the raw Detail Summary, 'sentiment_Vader', and the cleaned Detail Summary, 'sentiment_nltk'. Using the raw Detail Summary, a more balanced set is achieved suggesting the VADER text analyzer is using some of the features, like punctuation and capitalization, that is usually cleaned. These categorical values are used in the classification machine learning models described below.

Figure 15: Labeled Sentiment Values Distribution



5.2 Predictive Models using Regression

5.2.1 Sightings, Sentiment, Religion/Politics

Using Regression models, we sought to predict the sentiment and objectivity of sightings using available features, namely religious and political affiliation.

[5]https://medium.com/@rahulvaish/textblob-and-sentiment-analysis-python-a687e9fabe96

Feature Selection For sightings, there are factors in the sightings dataset to use to attempt to predict sentiment but those features do not address in any significant way the hypothesis. Therefore, we wanted to use the Pew Research religious beliefs and/or political affiliation. This data is summarized at the state level, so those factors were imputed to each sighting based upon the state in which the sighting was made.

As referenced above, there are five religious beliefs and/or political affiliations for two separate years for a total of twenty factors.

With twenty factors, we sought to reduce the dimensionality of the model. Using Principal Component Analysis (or PCA), we look for the features that cause the majority of the variance, using only those features thus reducing the complexity of the model.

PCA shows that seven features achieve a goal of 95% of the total variance: rel1_2007, rel1_2014, par2_2014, rel2_2007, rel3_2014, par3_2014, rel3_2007.

Model Training With our features reduced to those seven, we are ready to train a model. We started with a simple Linear Regression with the NLTK sentiment as the label. Eventually, we included multiple regressor models, all from SKLearn, to provide an improved model. The complete list of regression models used is below:

- Gradient Descent - SGDRegressor()
- Linear Regression - LinearRegression()
- Random Forest - RandomForestRegressor()
- K-nearest Neighbor - KNeighborsRegressor()
- Decision Tree - DecisionTreeRegressor()
- Support Vector Machine - SVR(kernel='rbf')

Model Evaluation For evaluation of the model, Root Mean Squared Error (RMSE) and R^2 are used. Ideally, RMSE should be as close to zero as possible as it represents the cumulative error while R^2 is a value between zero and one (in theory), with one being the best.

Figure 16: Regression Model Results - PCA

	Model	Training RMSE	Training R2	Test RMSE	Test R2
0	SGDRegressor()	0.244852	0.001203	0.245722	0.001561
1	LinearRegression()	0.244837	0.001265	0.245753	0.001436
4	DecisionTreeRegressor()	0.244314	0.003396	0.245806	0.001222
2	(DecisionTreeRegressor(max_features='auto', ra...	0.244319	0.003379	0.245817	0.001178
5	SVR()	0.262155	-0.069380	0.265533	-0.078934
3	KNeighborsRegressor()	0.308426	-0.258125	0.307381	-0.248973

Figure 17: Regression Model Results - Full

	Model	Training RMSE	Training R2	Test RMSE	Test R2
1	LinearRegression()	0.244722	0.001733	0.245666	0.001791
0	SGDRegressor()	0.244906	0.000984	0.245694	0.001677
2	(DecisionTreeRegressor(max_features='auto', ra...	0.244317	0.003387	0.245792	0.001277
4	DecisionTreeRegressor()	0.244314	0.003396	0.245806	0.001222
5	SVR()	0.262103	-0.069165	0.265487	-0.078750
3	KNeighborsRegressor()	0.274890	-0.121325	0.278705	-0.132459

The final models demonstrate that the dimensionality reduction does not adversely affect the model, but does improve performance. However, the model is very poor. Gradient Descent (PCA model) shows the "best" RMSE but with an R^2 value very low (0.001561) it does appear that we can predict the sentiment of a UFO sighting based upon the religious beliefs and/or political affiliations of the state.

5.2.2 Number of Sightings Per State

To further test the hypothesis that religious beliefs and/or political affiliation may affect UFO sightings, we sought to predict the number of sightings per state based on the same factors as the previous Regression analysis. In this case, the religious and political data originates at the state level, so it seems natural to use that state data to make predictions.

Feature Selection Similar to the Regression on the Sightings Sentiment predict, we wanted to reduce the dimensionality of the dataset. To gain the 95% variance goal nine factors are needed: rel1_2007, rel1_2014, rel2_2007, rel2_2014, par1_2007, par2_2007, par1_2014, par2_2014, rel0_2007.

Model Training and Evaluation The same algorithms were used again.

Figure 18: Regression Model Results - PCA

	Model	Training RMSE	Training R2	Test RMSE	Test R2
5	SVR()	1.825195e+06	-0.091038	2.398126e+05	-0.306994
3	KNeighborsRegressor(n_neighbors=2)	4.510488e+05	0.730379	5.459628e+05	-1.975534
0	SGDRegressor(max_iter=500000)	6.138788e+05	0.633045	9.578175e+05	-4.220169
1	LinearRegression()	6.096644e+05	0.635564	1.006255e+06	-4.484154
2	(DecisionTreeRegressor(max_features='auto', ra...	2.270701e+05	0.864265	1.694018e+06	-8.232511
4	DecisionTreeRegressor()	0.000000e+00	1.000000	6.261972e+06	-33.128166

Figure 19: Regression Model Results - Full

	Model	Training RMSE	Training R2	Test RMSE	Test R2
5	SVR()	1.825237e+06	-0.091063	240007.338558	-0.308056
2	(DecisionTreeRegressor(max_features='auto', ra...	2.224467e+05	0.867029	524157.049965	-1.856691
3	KNeighborsRegressor(n_neighbors=2)	4.554872e+05	0.727726	550494.075000	-2.000229
4	DecisionTreeRegressor()	0.000000e+00	1.000000	577165.000000	-2.145588
0	SGDRegressor(max_iter=500000)	2.803200e+05	0.832434	679049.117546	-2.700863
1	LinearRegression()	2.340911e+05	0.860068	690922.000000	-2.765571

With the exception of the Support Vector Machine on PCA data, the other models looked very promising in training, however when predicting using Test data we discover a breathtaking overfitting.

5.3 Predictive Models using Classification

Since the regression models did not show promising results with a continuous output, we decided to model the categorical labels that were created earlier by bucketizing the sentiment values into 'very positive', 'positive', 'neutral', 'negative', and 'very negative'. The models trained and evaluated in this analysis were:

- Multinomial Naive Bayes Classifier
- KNeighbors Classifier
- Logistic Regression
- MLP Classifier
- Support Vector Classifier

Feature Selection Based on the dimensional reduction described above with PCA, the classification model training included only the PCA dataset since the regression models performed similarly with the full and reduced datasets.

Model Training and Evaluation Figures 20 and 21 show the confusion matrices , training, and test accuracy for each model. The accuracy maxed out at 47% with most of the models showing this accuracy for both train and test datasets. After inspection of the value counts, the dataset was found to be severely un-balanced with nearly half of the data with a label of 4/ or 'very positive'. The dataset was then balanced using the SMOTE, Synthetic Minority Oversampling TEchnique, with the PCA datasets and re-fit to the same classifiers. The accuracy dropped to around 20% for all models indicating the predictive power of the features are low.

Figure 20: Classification Model Results - Full

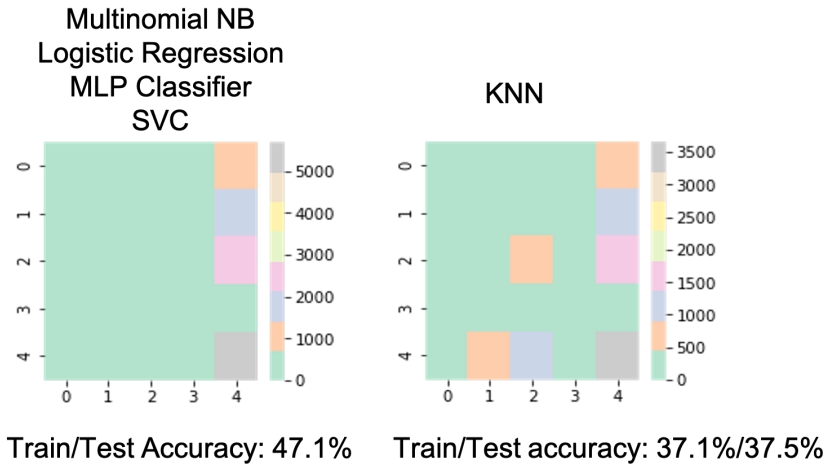
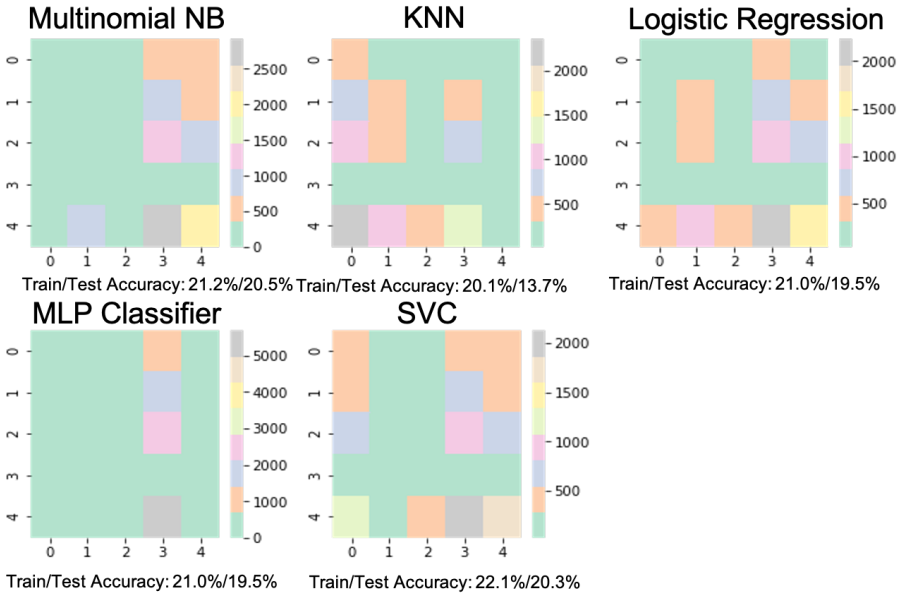


Figure 21: Classification Model Results - PCA Balanced



6 Conclusions

Sentiment analysis is one of the fastest growing research areas in data science, and a lot of improvements have been made recently. One of the goals of this analysis was to determine the best sentiment analyzer to use with UFO sighting reports. We argue that the VADER sentiment analyzer using the un-cleaned UFO reports is the best since it provided a slightly wider range of values. This is due to the recent additions into the algorithm that utilizes capitalization and punctuation in the sentiment parameter calculation. Based on this observation, it is recommended to use the VADER sentiment analyzer using the un-cleaned text in the case of UFO sighting reports.

Another goal of this analysis was to determine if a relationship exists between the sentiment of UFO sighting reports and the religious or political affiliation in the area.

Religious data between the years of 2007 and 2014 showed most people in the U.S. identify as Protestant. Mormonism was the number one religion in Utah in both 2007 and 2014, while in Massachusetts, Rhode Island, it was number 1 in 2007, then dropped in favor of Judiasm in 2014.

Within the political data, it was observed that the number of people that responded 'No preference' or 'Don't know' was relatively large in 2007 while the number was considerably smaller in 2014. This observation suggests people are more willing to identify with one of the 'big' parties, Democrat or Republican.

Using this UFO sightings dataset and this data, we argue there is no correlation between religious beliefs, political leanings and UFO sightings data, whether that is the sentiment or objectivity of the sighting or the number of sightings by region.

Another trend observed was in the count trends for number of reports seen by year, day and hour. Between the years of 2006 and 2020, the highest number of sightings were seen in 2014 and 2020. The most common days for sightings are on the 4th of July and Dec 21. The hours most of the sighting occur is between 4 and 5 in the morning.

Finally, some interesting trends in the data were observed. Using geographic data and mapping, as shown above, we can see that the Southwest United States does gather sightings more than any other region. The Northwest and Southeast has some increased activity. However, the Southwest, including Southern California, Arizona, and Nevada, have the most activity. Further research may be required to determine a cause. The data in this project does not indicate an explanation for the high number of sightings in these areas.