Network Analytics Project Report

Sarcasm Detection in News Headlines

Submitted by: Team Jarvis

Aishwarya Tiwari, Divisha Jain, Naveen Parathasarathy, Priyanka Demla

Understanding the Dataset:

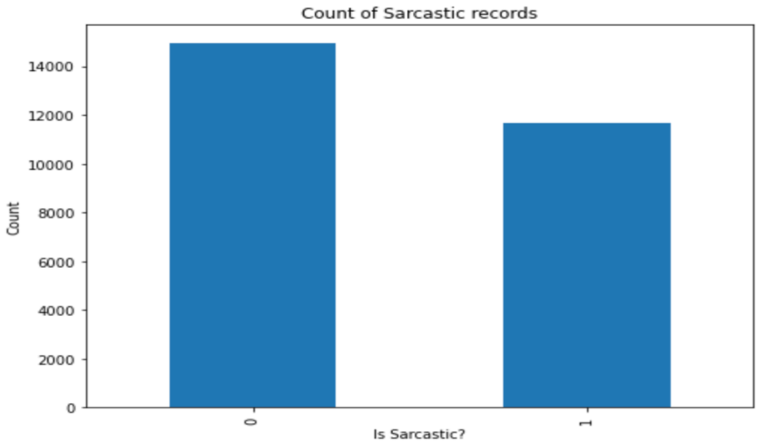
We have used a labelled dataset from Kaggle. The labeled dataset focuses on news headlines extracted from two distinct newspapers, namely TheOnion and HuffPost. Notably, content from TheOnion is characterized by its satirical perspective on current events, offering a sarcastic commentary. The dataset encompasses information spanning the "News in Brief" and "News in Photos" categories, contributing a unique dimension to the compilation of data for analysis and exploration. The dataset has 28,619 entries which gave us a decent sample size to work with.

Pre-Processing:

Before the Exploratory data analysis, we did some NLP pre-processing to transform the data, which would ensure that our EDA makes sense. We performed keyword extraction from text data using the NLTK for stopwords & punctuations, lemmatization on the 'headline' column using the Word class from the textblob library and Pandas for data handling.

Exploratory Data Analysis:

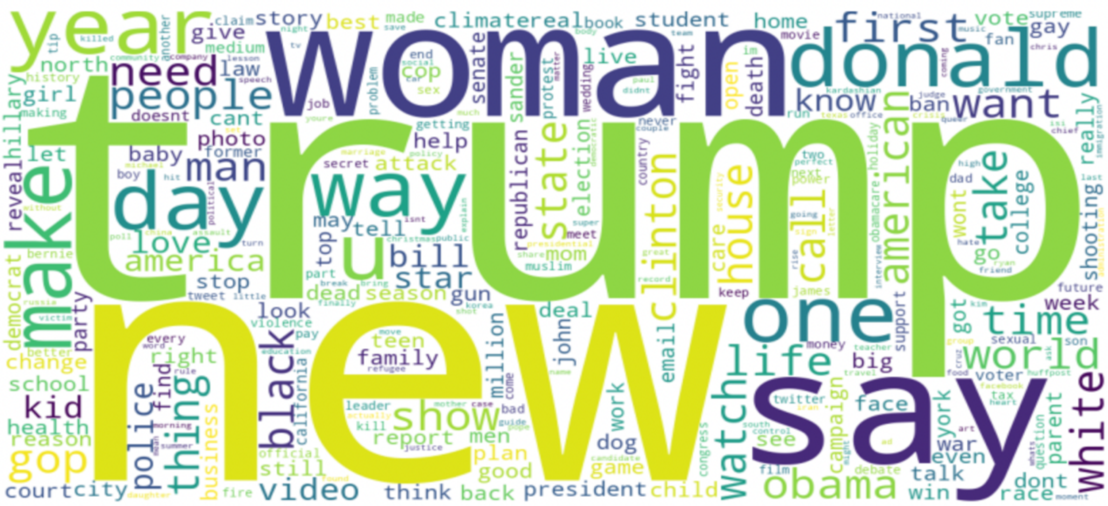
For the Exploratory Data Analysis, we use the Pandas/Numpy, seaborn, Matplotlib and WordCloud libraries in Python to create visualisations to explore the dataset.



Since the distribution of sarcastic and serious headlines is almost equal, our dataset is balanced and we can proceed.

Then we created two different wordclouds.

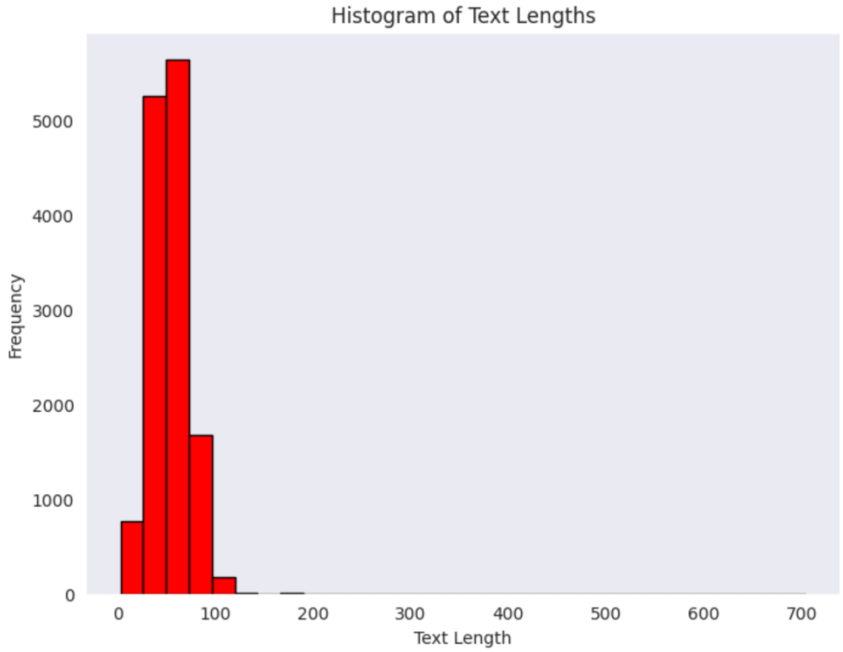
First, we created a word cloud from the headlines of non-sarcastic articles in the DataFrame and displayed it using Matplotlib. The word cloud visually represents the most frequent words in the non-sarcastic headlines, with the size of each word indicating its frequency.



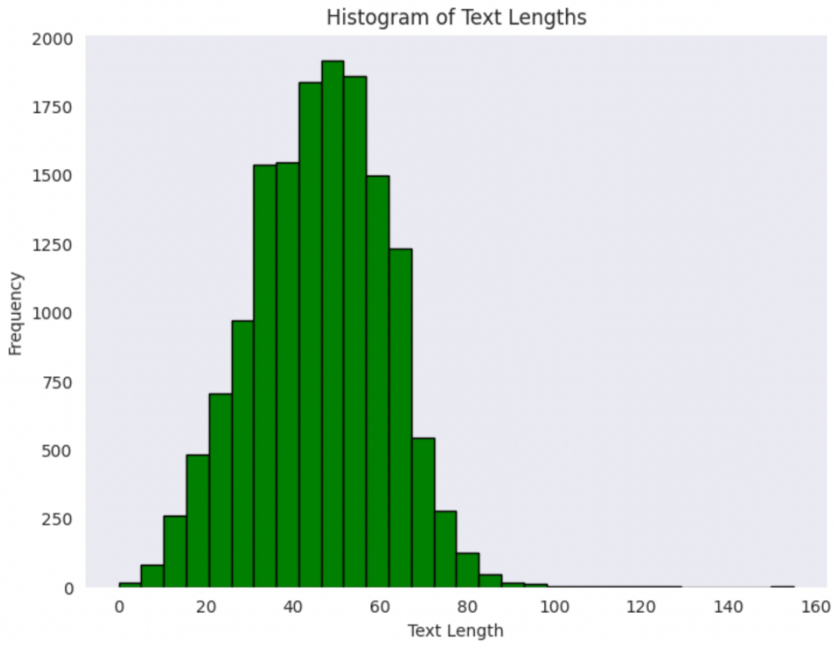
Next, we created a similar word cloud from the headlines of sarcastic articles in the DataFrame!

Plotting histograms comparing text lengths between sarcastic and non-sarcastic articles:

```
Minimum Length: 3
Maximum Length: 704
25th Percentile: 40.0
50th Percentile (Median): 52.0
75th Percentile: 65.0
99th Percentile: 100.0
```

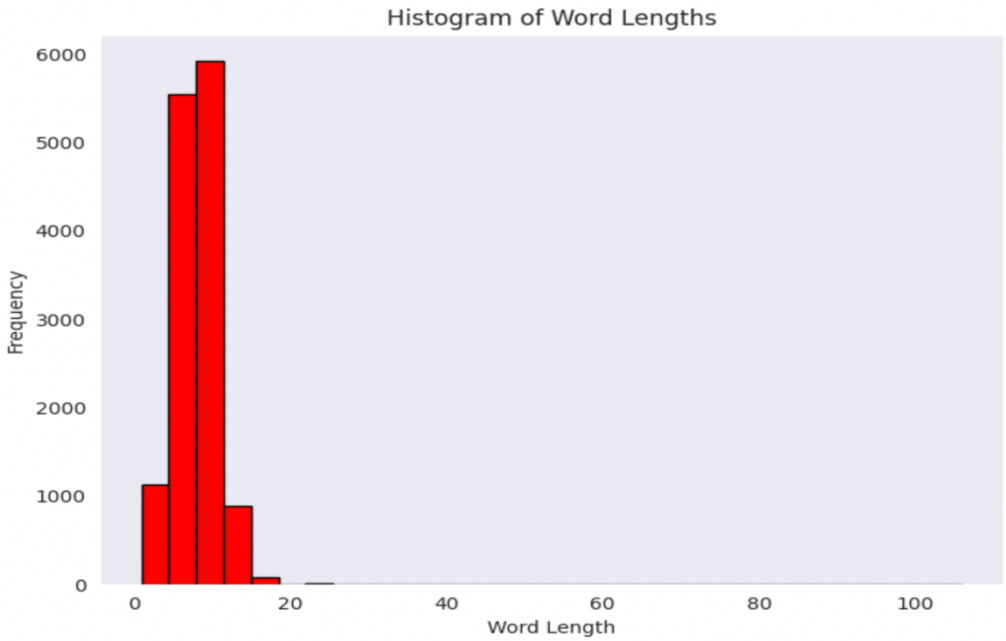


```
Minimum Length: 0
Maximum Length: 155
25th Percentile: 35.0
50th Percentile (Median): 47.0
75th Percentile: 57.0
99th Percentile: 80.0
```

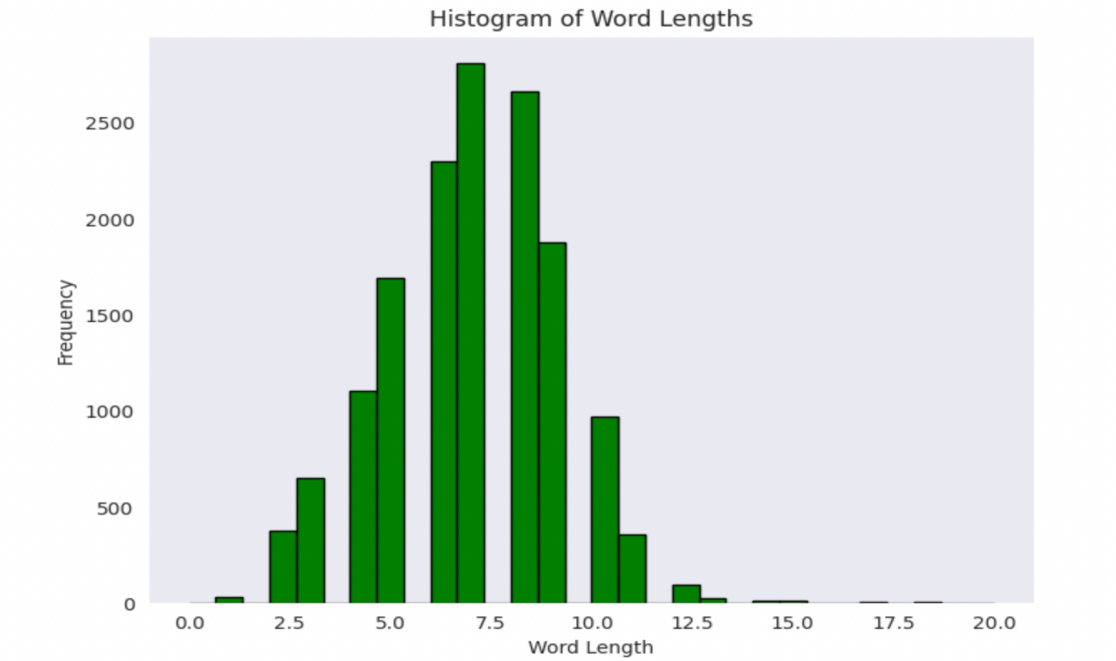


Plotting histograms comparing word lengths between sarcastic and non-sarcastic articles:

```
Minimum Length: 1
Maximum Length: 106
25th Percentile: 6.0
50th Percentile (Median): 8.0
75th Percentile: 9.0
99th Percentile: 14.0
```



```
Minimum Length: 0
Maximum Length: 20
25th Percentile: 5.0
50th Percentile (Median): 7.0
75th Percentile: 8.0
99th Percentile: 12.0
```

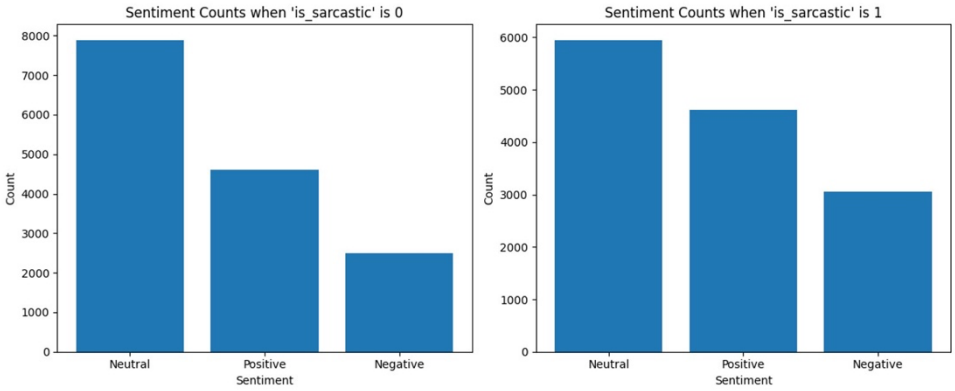


Insights from EDA:

In examining the distribution of both the number of words and the length of individual words, we observed an approximate normal distribution. Further analysis revealed a notable trend where sarcastic headlines exhibited a tendency to be longer compared to their serious counterparts. Notably, the longest serious headline in the dataset comprised only 150 words. However, the presence of outliers was also identified among certain headlines. In light of this, it is proposed to exclude these outliers from the dataset, particularly in preparation for topic modeling, to ensure a more focused and representative analysis.

Sentiment Analysis:

The results from Sentiment Analysis are as follows:



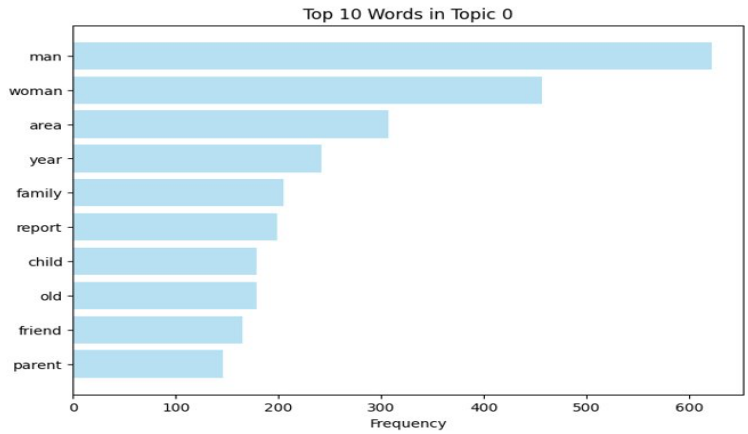
For non-sarcastic headlines:

The prevailing sentiment classification indicates that a substantial majority of news headlines have been categorized as possessing a neutral sentiment. Following closely are a smaller proportion characterized by positive sentiments, while the least number fall into the category of negative sentiments. This observation suggests a prevailing trend where the majority of authentic news headlines are crafted with an emphasis on delivering information rather than evoking emotional responses.

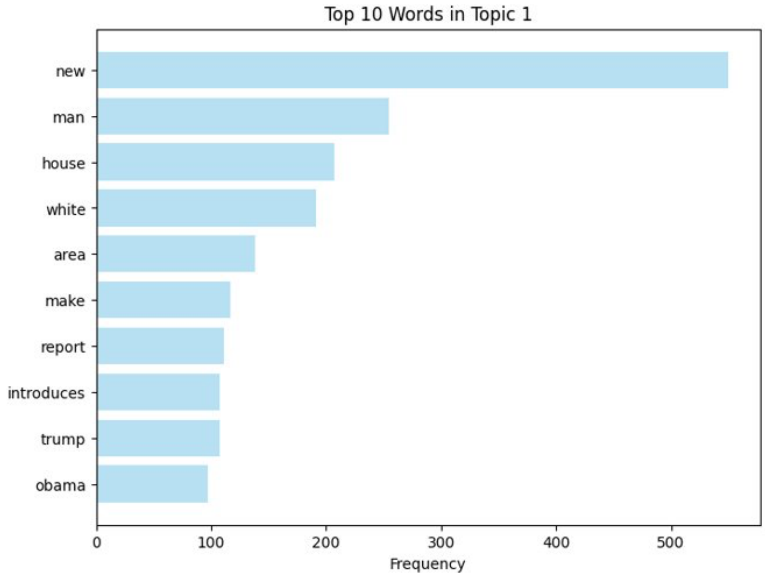
For sarcastic headlines:

While the prevalence of neutral sentiment remains dominant, there is a noteworthy proximity between the count of positive sentiment headlines and the neutral count. This observation suggests a potential pattern in sarcastic headlines, wherein they tend to emulate a subtly positive tone. This aligns with a common characteristic of sarcasm, where positive language is employed to convey a negative underlying meaning.

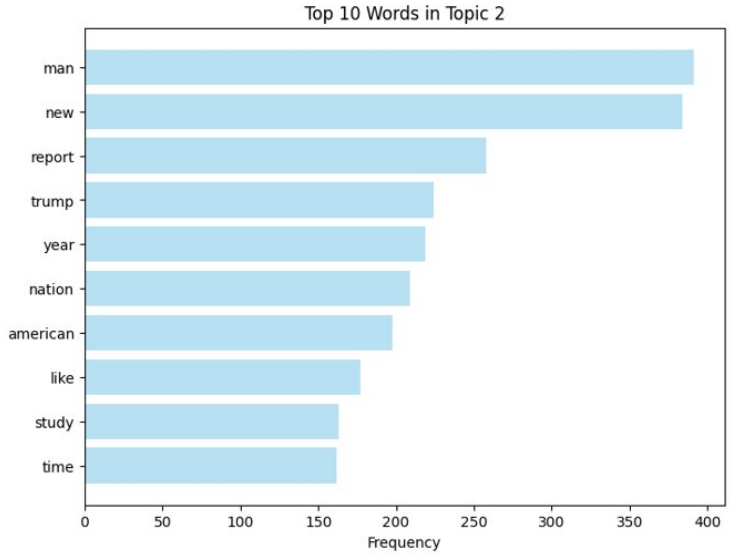
Topic Modelling: Sarcastic Headlines



Topic 0 revolves around keywords such as man, woman, area, year, family, report, child, old, friend, and parent. The thematic focus encompasses wellness, living, relationships, and family-related subjects. In the context of sarcastic implementation, this topic involves addressing the routine, mundane, or seemingly ordinary aspects of life infused with a sarcastic twist.

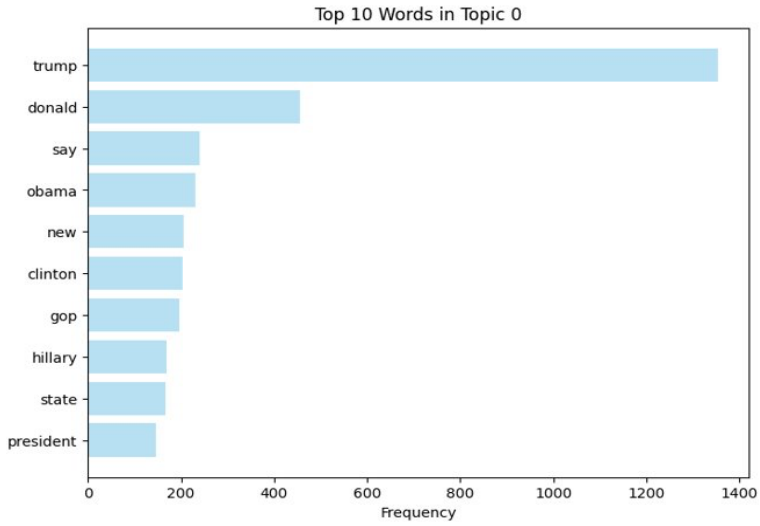


Topic 1 is characterized by keywords such as new, man, house, white, area, make, report, introduces, Trump, and Obama. The thematic focus includes political figures, events, and policies, with a particular emphasis on housing-related issues. Additionally, this topic addresses any new legislation that may be introduced. In the context of sarcastic commentary, it involves a satirical take on political developments, policy changes, or news reports within the specified domain.

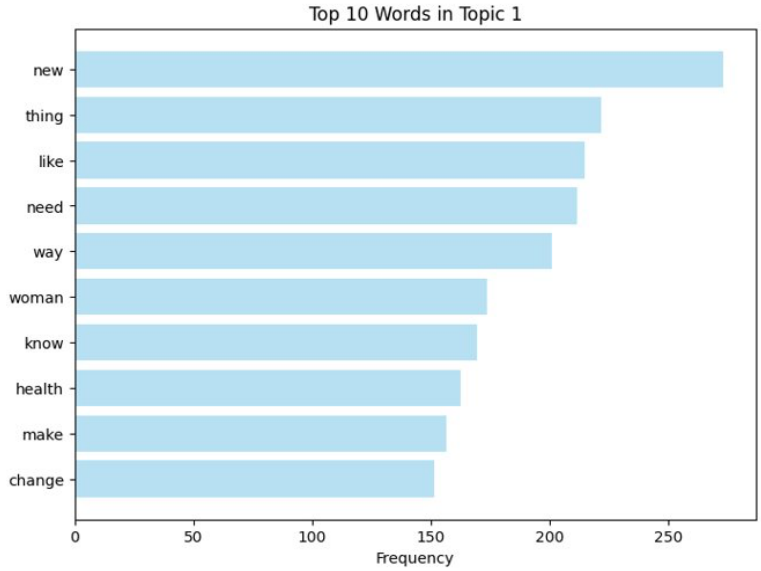


Topic 2 is characterized by keywords such as man, new, report, Trump, year, nation, American, like, study, and time. The focus of this topic centers around American politics, with an emphasis on a national perspective rather than addressing specific issues. The discussion within this topic encompasses broader themes within the realm of American political dynamics.

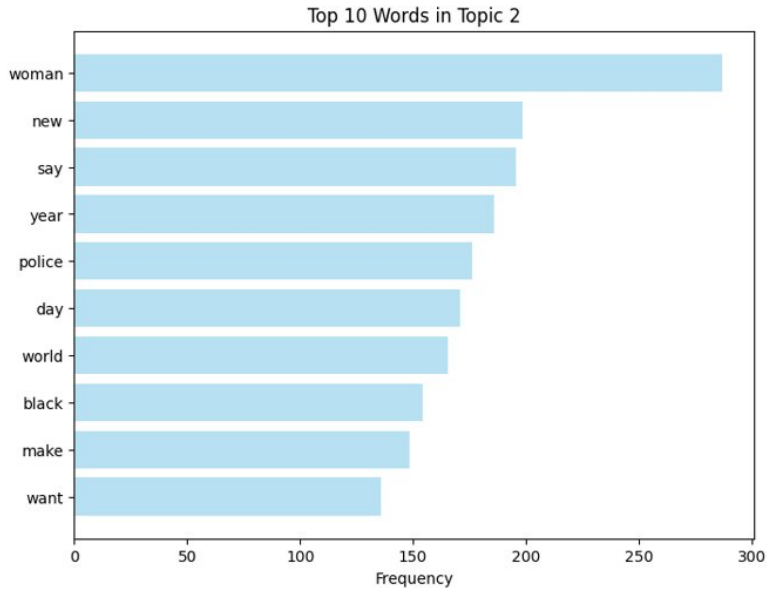
Topic Modelling: Serious Headlines



Topic 0 is characterized by keywords such as Trump, Donald, say, Obama, new, Clinton, GOP, Hillary, state, and president. This topic delves into the landscape of American politics and presidential administrations, highlighting the names of prominent political figures associated with the discourse. The thematic focus revolves around discussions related to key individuals within the political sphere.



Topic 1 is characterized by keywords such as new, thing, like, need, way, woman, know, health, make, and change. The primary thematic elements encompass lifestyle, health, and overall well-being. The presence of words such as "like," "need," "way," and "know" suggests a focus on informative or advisory content within this topic. The discussions within this context are likely centered around guiding and informing the audience on various aspects related to lifestyle and health.

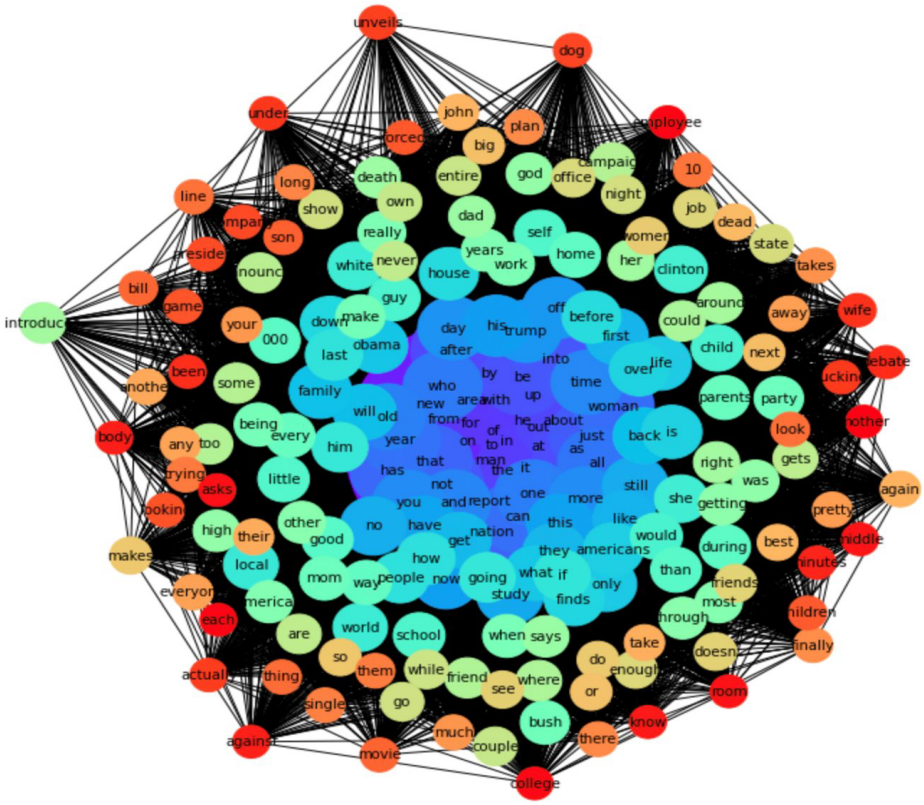


Topic 2 is characterized by keywords such as woman, new, say, year, police, day, world, black, make, and want. This topic suggests a potential emphasis on social issues, with keywords like "woman," "black," and "police" serving as indicators of stories related to social justice, civil rights, or matters pertaining

to policing. The thematic focus within this context is likely to explore narratives that touch upon significant societal and global concerns.

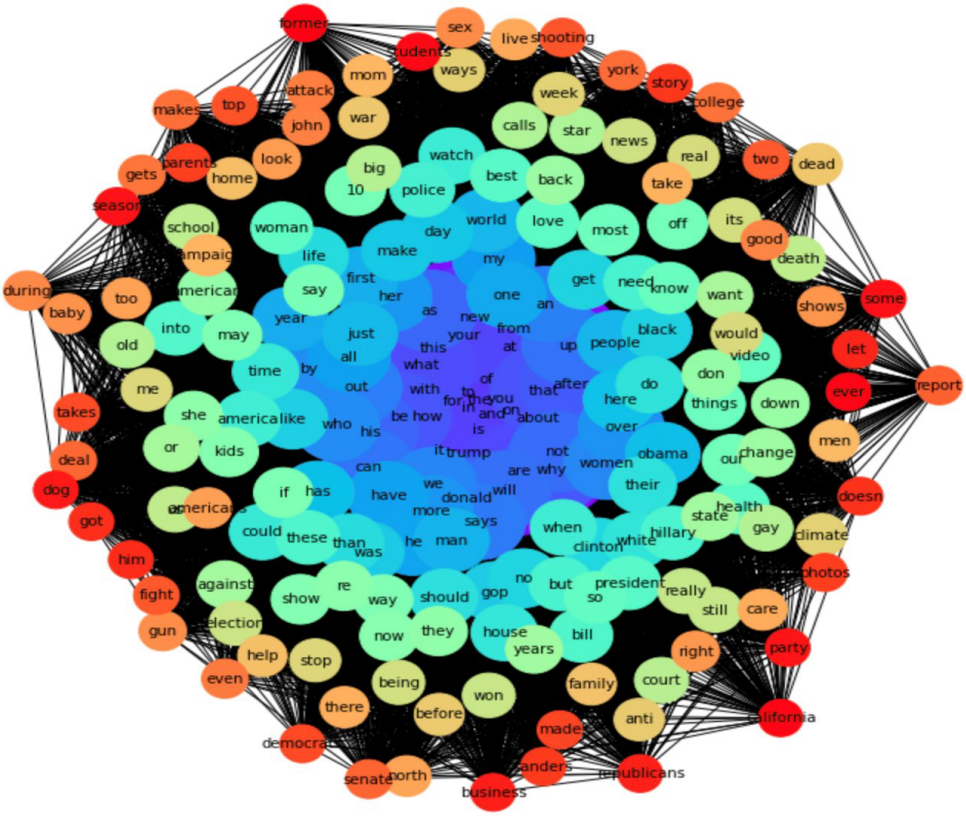
Network Visualization:

Sarcastic Headlines -



The analysis of sarcastic headlines reveals three distinct topics. Firstly, the political theme is prevalent, featuring keywords such as Trump, Obama, Clinton, and state, indicating a focus on American politics and prominent political figures. The second topic centers around family-related content, with emphasis on words like dog, wife, children, and mother. This theme likely explores sarcastic commentary on familial dynamics. Lastly, the third topic encompasses fun activities, incorporating keywords like college, movie, and game. In this category, sarcastic headlines appear to playfully address leisure and recreational aspects, contributing a diverse range of satirical perspectives across politics, family, and leisurely pursuits.

Non Sarcastic Headlines -



The observed clusters in the analysis align with conventional news categories, encompassing topics such as "health care," "climate," and "election." These clusters indicate a structured organization within the dataset, suggesting that the headlines tend to gravitate toward standard news themes and subject matter commonly found in news reporting, providing a basis for further exploration and understanding of the underlying patterns in the dataset.

Overall, in both graphical representations, central nodes stand out as probable key subjects or recurring terms within the headlines. Notably, terms such as "election," "America," "president," and "Obama" emerge prominently, suggesting their significance as pivotal elements within the dataset. These findings indicate a thematic focus on political discourse, with a particular emphasis on elections, the United States, and notable political figures such as the president. The visual representation of these key terms provides valuable insights into the prevalent subjects and recurring themes across the analyzed headlines.

Classification Model:

We tried evaluating the dataset using classification model based on the following two methods:

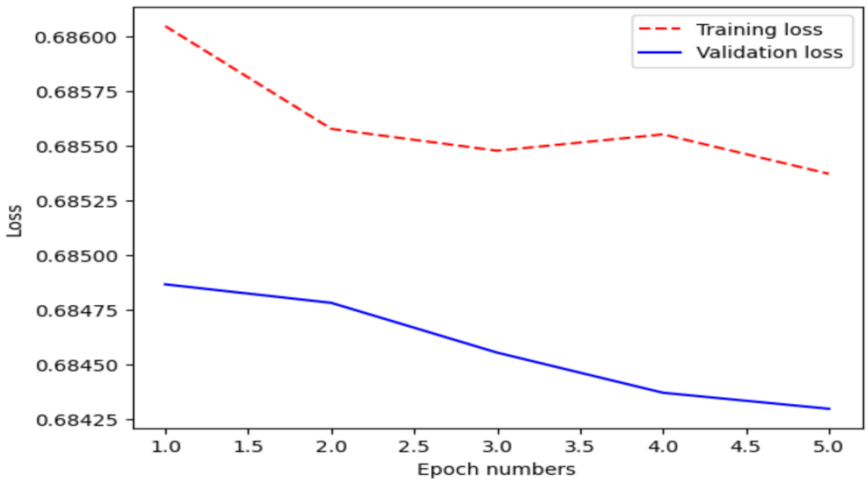
Long Short-Term Memory (LSTM) Method:

The LSTM (Long Short-Term Memory) layer is a type of recurrent neural network (RNN) layer designed to capture and learn long-term dependencies in sequential data. We first tried to build a neural network model for binary classification, with an embedding layer, an LSTM layer.

The model summary is as follows:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 229, 100)          5100

 lstm (LSTM)                 (None, 229, 64)           42240

 dense (Dense)               (None, 229, 25)           1625

 dropout (Dropout)           (None, 229, 25)           0

 dense_1 (Dense)             (None, 229, 1)            26

=================================================================
Total params: 48991 (191.37 KB)
Trainable params: 48991 (191.37 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

The following line plot shows the training and validation loss over epochs, making it easy to visualize how well the model is learning from the training data and generalizing to unseen validation data.



As seen above, the accuracy for the model was pretty low so we tried to use another model.

Bi Directional Long Short-Term Memory (LSTM) Method:

Bidirectional LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that processes sequential data in both forward and backward directions. It captures contextual information from past and future inputs, enhancing understanding of sequences and context in tasks like text generation, translation, and sentiment analysis. By leveraging information from both directions, Bidirectional LSTMs excel in learning long-range dependencies, crucial for understanding complex sequences in natural language processing and other sequential data tasks.

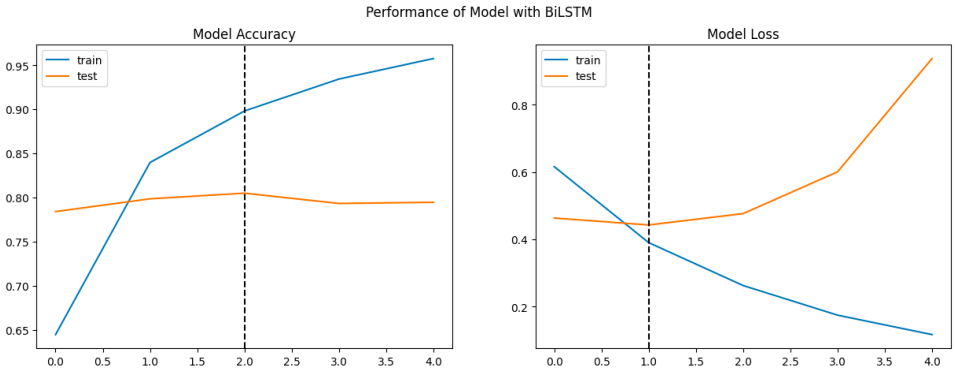
Model Summary:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 200)         2000000

 bidirectional (Bidirection  (None, None, 256)         337920
 al)

 global_max_pooling1d (Glob  (None, 256)               0
 alMaxPooling1D)

 dense (Dense)               (None, 40)                10280

 dropout (Dropout)           (None, 40)                0

 dense_1 (Dense)             (None, 20)                820

 dropout_1 (Dropout)         (None, 20)                0

 dense_2 (Dense)             (None, 1)                 21

=================================================================
```

Training and Validation accuracies are below -

```
Epoch 1/5
213/213 [==============================] - 26s 77ms/step - loss: 0.6159 - accuracy: 0.6445 - val_loss: 0.4635 - val_accuracy: 0.7841
Epoch 2/5
213/213 [==============================] - 8s 36ms/step - loss: 0.3905 - accuracy: 0.8398 - val_loss: 0.4428 - val_accuracy: 0.7985
Epoch 3/5
213/213 [==============================] - 3s 13ms/step - loss: 0.2631 - accuracy: 0.8982 - val_loss: 0.4766 - val_accuracy: 0.8049
Epoch 4/5
213/213 [==============================] - 3s 16ms/step - loss: 0.1751 - accuracy: 0.9344 - val_loss: 0.6007 - val_accuracy: 0.7933
Epoch 5/5
213/213 [==============================] - 4s 20ms/step - loss: 0.1173 - accuracy: 0.9577 - val_loss: 0.9370 - val_accuracy: 0.7946
```

Model performance is better but still not at a high level.



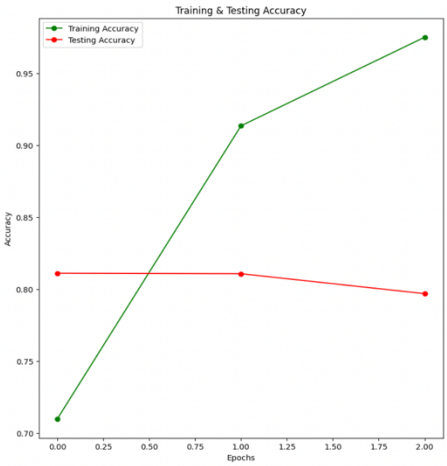
Word2Vec Method:

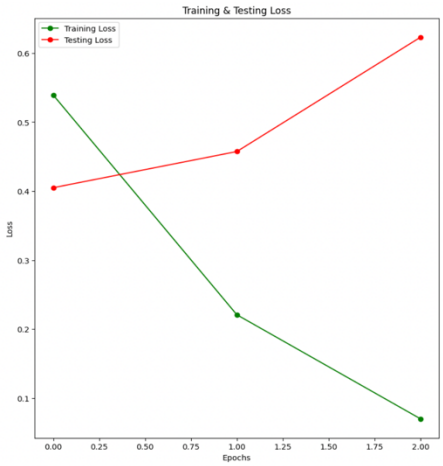
Word2Vec is a neural network model that learns distributed representations (word vectors) of words in a continuous vector space from large amounts of unlabeled text data.

The model summary after we used the Word2Vec Method was as follows:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 20, 200)           7614400

 bidirectional (Bidirection  (None, 20, 256)           336896
 al)

 bidirectional_1 (Bidirecti  (None, 64)                55680
 onal)

 dense (Dense)               (None, 1)                 65

=================================================================
Total params: 8007041 (30.54 MB)
Trainable params: 8007041 (30.54 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

We observed that, while the model achieved a high accuracy of 99% on the training dataset, its accuracy on the testing dataset dropped to 80% so there is a clear case of overfitting on training data that we found!





Finally we took the predicted probabilities from the neural network model, applied a threshold of 0.5 and converted the probabilities into binary predictions. This is a common post-processing step when dealing with binary classification tasks.

BERT Method:

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art natural language processing (NLP) model developed by Google. It uses bidirectional context to understand word meaning within sentences, capturing deeper language nuances. By pre-training on large text corpora, BERT learns rich representations, enabling it to excel in various NLP tasks like question answering, text classification, and more.

```
 Layer (type)                Output Shape              Param #
=================================================================
 input_word_ids (InputLayer  [(None, 16)]              0
 )

 tf_bert_model (TFBertModel  TFBaseModelOutputWithPo   109482240
 )                           olingAndCrossAttentions
                             (last_hidden_state=(Non
                             e, 16, 768),
                              pooler_output=(None, 7
                             68),
                              past_key_values=None,
                             hidden_states=None, att
                             entions=None, cross_att
                             entions=None)

 lambda (Lambda)             (None, 768)               0

 dense_1 (Dense)             (None, 128)               98432

 dropout_37 (Dropout)        (None, 128)               0

 dense_2 (Dense)             (None, 1)                 129
...
Total params: 109580801 (418.02 MB)
Trainable params: 109580801 (418.02 MB)
Non-trainable params: 0 (0.00 Byte)
```

We observed that, while the model achieved a high accuracy of 99% on the training dataset, its accuracy on the testing dataset was 87%, which is better than the Word2Vec model.