**BA820 B1 Team 06**

Kevin Murphy, Khushi Manish Jasrapuria, Megha Arul Senthilkumar, Riris Grace Karolina, Shravani Thalla

**Project title: Beyond the Headlines: Machine Learning Insights for News Articles**

**Github Link:** [**https://github.com/kevin-r-murphy/ba820/tree/main**](https://github.com/kevin-r-murphy/ba820/tree/main)

1. **Exploratory Data Analysis** **(EDA):**

Our data is sourced from the news headlines dataset containing data from 2012 to 2022 from [HuffPos](https://www.huffingtonpost.com/)t available on Kaggle: <https://www.kaggle.com/datasets/rmisra/news-category-dataset>. This dataset was initially provided in a JSON file format. Based on the EDA, we observed the following:

* Identified 209,527 rows and 6 columns: link, headline, short description, authors, date, and category (which also serves as the label).
* Our data is evenly distributed from 2012 to 2017 but shows a decline from 2018 to 2022. Also, analyzed the distribution of character length for headlines over years.
* The word 'Trump' appears most frequently in the news from 2016 to 2022 which supports “Politics” being the most common category in the dataset, accounting for approximately 18% of the news.

1. **Preprocessing:**

While our dataset was generally clean, there were a few rows that needed to be excluded because they contained empty strings. To standardize the text, we converted all characters to lowercase, eliminated punctuation and spaces, removed non-English characters and applied lemmatization before conducting our analysis.

1. **Analysis Plan:**

Our primary objectives were to :

* Predict the news category based on the headlines and short descriptions

The initial plan was to use classification and clustering to classify news article headlines and short descriptions. Since the categorization was already present in the dataset, we planned on using classification to build a model that predicts categories and clustering techniques to explore any patterns that can be found.

* Analyze whether the news article is providing positive or negative news.

The initial plan was to use sentiment analysis to classify the headlines and short descriptions into positive and negative news categories. As this classification was not available in the dataset we planned on using unsupervised machine learning to identify positive and negative articles.

1. **Final Results:**To address the objectives mentioned above, we conducted several pre-processing steps including data cleansing and tokenizing, then ran several models.

Our project can be divided based on its to objectives as follows:

1. Predicting the category of news based on the Headline and Short Description

* **Bag of Words:** Leveraged the BOW model to tokenize and vectorize. Further, we used inverse transform to transform them back to their original textual format.
* **Document similarity:**Utilized cosine similarity to identify similar headlines to given words.
* **TF-IDF & TF-IDF Weighted Average:**Leveraged the TF-IDF model to tokenize, vectorize and classify based on it.
* **Basic Classification Model:**Used the basic Logistic Regression model on the tokens and vectors generated to check for accuracy before we moved on to use more sophisticated models.
* **N-grams Model:**Classifying the news using an n-gram model (ranging from unigrams to trigrams) by training on and splitting the data.
* **Word2Vec Model & Google news’s Word2Vec Model:**Leveraged the Word2Vec and Google news’s Word2Vec models to generate embeddings and used them to classify the headlines, short descriptions and both together using Logistic Regression, Decision Tree and Random Forest Classification.

Based on our analysis, the BOW Model produced less significant results than the TF/IDF, and the N-grams model produced time constraints.thus no additional analysis was conducted using it.

* **Clustering**: Done k-means clustering for the combined headline and short description for it to provide the natural groups. Cluster 1 had more Lifestyle and leisure categories . Cluster 2 had more of well-being and family and less of politics. Cluster 3 had politics and news-related categories. (Plot in Appendix)

The results from our final model are as follows:

| Model | Headline | Short description | Combination of Headline and Short Description |
| --- | --- | --- | --- |
| Word2Vec (build own model) | 50.9% | 40.9% | 54.5% |
| Word2Vec (using Google News Model) | 46.8% | 38.5% | 50.7% |
| Word2Vec TF-IDF Weighted Average (Own Model) | N/A | N/A | 54.6% |

Based on the models, we find that the best model is our own Word2Vec models trained on both the headlines and descriptions, with TF-IDF weighted average embeddings. While this was our best model, it is only slightly better than the non weighted average. At a large scale, where efficiency is important, the TF-IDF weighted average may be overkill. It is also important to note that our model only slightly outperformed the pretrained model, so it also may not be necessary to put the effort of training the Word2Vec model on the data.

1. Sentiment Analysis Results:

We used the Word2Vec module to analyze short descriptions rather than the headline as the sentences in the headline are fewer and tend not to accurately represent the tone of the news. Here are some results based on our observations:

* **Parameter adjustment**: we tried modifying several parameters, such as window, min\_count, and vector size to find the optimal for our model. The ones that work best are window = 6, min\_count = 1, and vector\_size = 60.
* **Similarity test & labeling**:Found that the model is better at recognizing negative sentiments based on the similarity test. Despite this, the model still correctly predicts the sentiment for each test sentence, as reflected by the similarity scores. After adding the new label for sentiment label on our data, we identify 57% of the news as having a positive tone, and 43% negative.
* **Categorical Check**: We checked to see the ratio of positive and negative headlines for each headline category(Chart in Appendix). We found that Political headlines made up the majority of the negative headlines.

1. **Challenges:**

* Dataset size : The dataset posed computational challenges for certain models. Initially, we employed a sample for model testing, later using the entire dataset to optimize parameters omitting bag of words, n-grams, and cosine similarity models during classification.
* Non-english words: In the preprocessing phase removed non-English words
* Empty Strings : While cleaning the data, we observed empty strings in specific rows. To avoid false assumptions, we removed the rows containing empty strings
* Class Imbalance:The number of articles in each category showed a notable difference which could have affected our model accuracy. However, recognizing that this imbalance reflects the real-world scenario, we chose not to intervene.
* Combining separate text embeddings into one model: We worked with two different text data: Headlines and Descriptions. We had to train Word2Vec models on each source separately, then combine the embeddings, before running a classification model on them

1. **Real world applications:**

Our headline classification model could be used by a news agency to automatically classify headlines and streamline their content management process, saving time and efforts. We believe that the real value could be driven from the sentiment analysis that was derived from the headlines as news agencies can check the tone of headlines before they are published and check that the headline matches the article before publishing and could also take the macro view of its headline sentiment, for example if its skewing negative over time, then the agency could focus on positive headline categories. Can give users a more personalized and relevant experience where readers are recommended topics they are interested in, causing more reader satisfaction and engagement.Content can be prioritized through clustering algorithms based on its perceived importance and can be given priority placement on news websites or mobile apps.

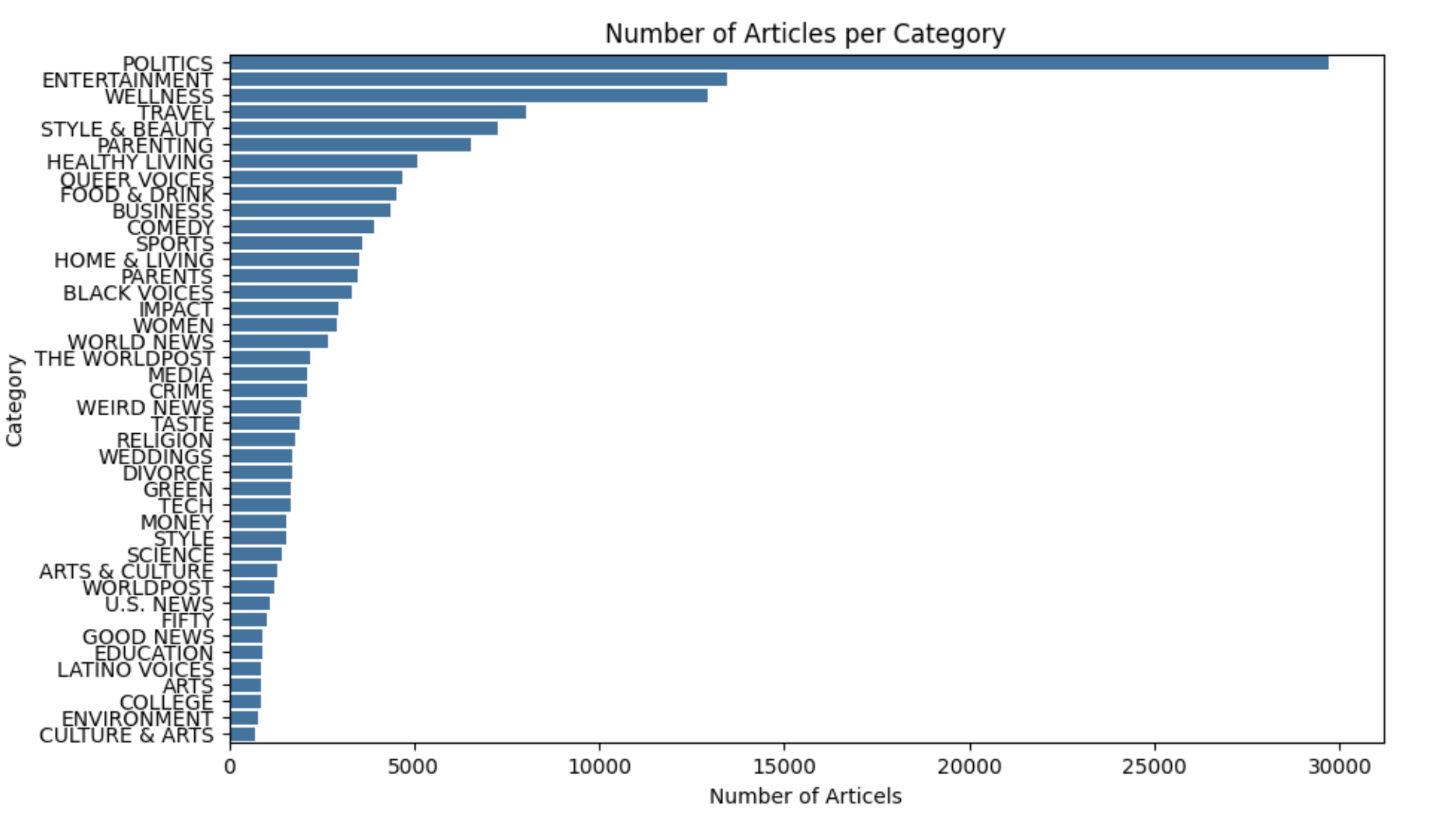
**APPENDIX**

GitHub Link:<https://github.com/kevin-r-murphy/ba820/blob/main/BA820.ipynb>

Contribution Sheet

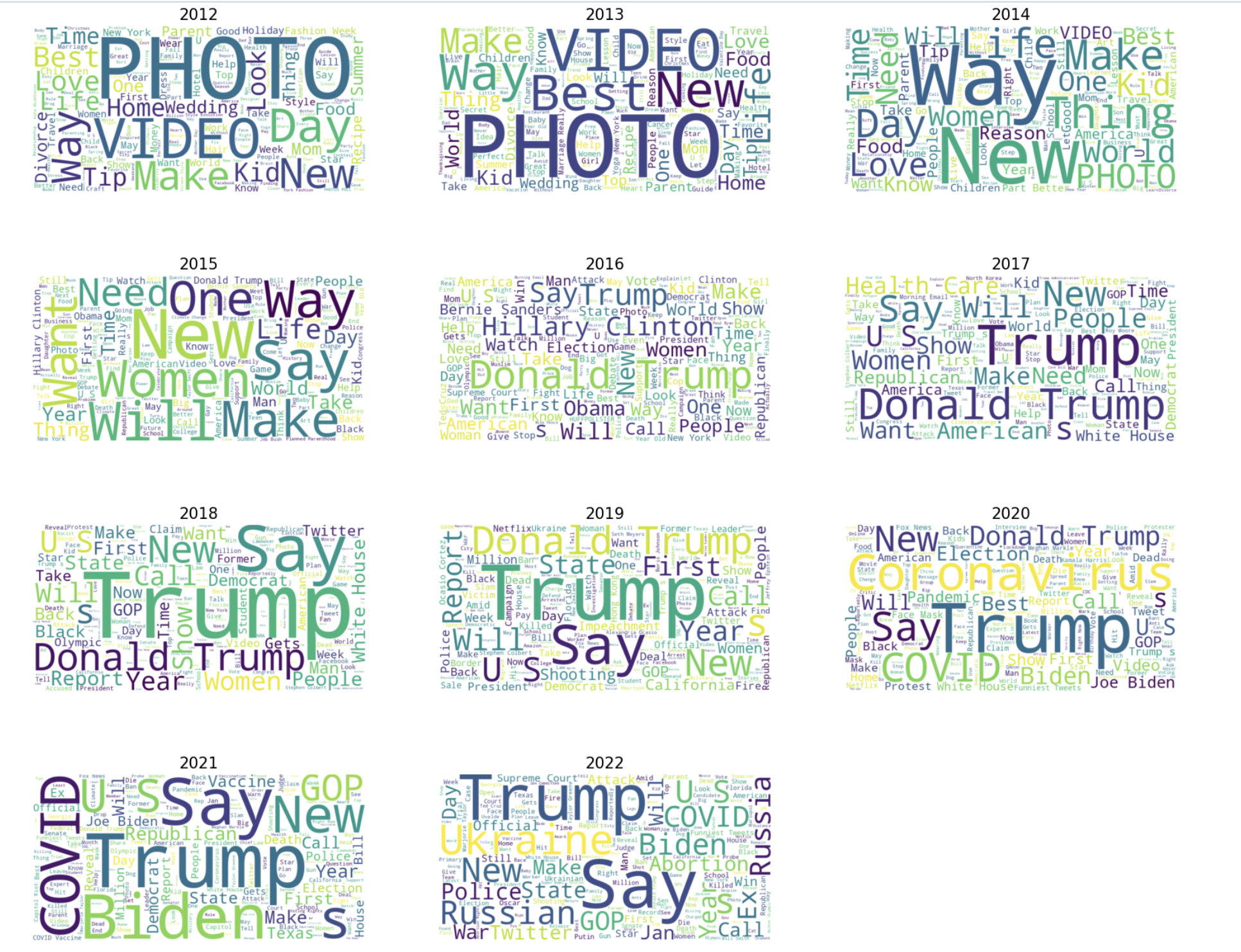
| **Member** | **Contribution** |
| --- | --- |
| Kevin | Notebook:   * Function to Stem, Lemm, and Tokenize * Word2Vec Pipeline * TF-IDF Weighted Average Pipeline * Helped with fixing Sentiment Analysis   + Stemm & Lemmed the ideal positive and negative headlines * Also in charge of running models on entire dataset (Colab Pro) |
| Shravani | Notebook:   * Data Source * Data Cleaning and data Inspection * EDA- created charts, word clouds and analysis * Part of Prepossessing or cleaning the text * Clustering ,PCA * Commenting on codes * Analysis   Report:   * EDA * Final Results |
| Khushi | Notebook:   * EDA - Preprocessing, created graphs and wrote analysis * Preliminary analysis including objectives, findings, metrics for evaluation and approaches for optimality & appendix * Modeling * TF-IDF * PCA * Clustering * Analysis * Commenting on codes * Results   Report:   * Final Results * Real World Applications |
| Riris | Notebook:   * Data Cleaning * Apply the Bag of Words * Classifying the news using an n-gram model * Evaluate the accuracy of both models using Supervised ML. * Sentiment Analysis: Apply Word2Vec model to identify the positive and negative sentiment, word embedding, testing the model with similarity testing, and apply the sentiment label to the dataset.   Report:   * EDA * Final Results * Real world application. |
| Megha | Notebook:   * Data Cleaning * Document Similarity * Word2Vec Model * Google news Word2vec Model * Classification models-logistic regression, decision tree random forest * Analysis   Report:   * Pre-Processing * Analysis Plan * Final Results * Challenges |

**Category Graph:**



The graph above shows the frequency of the number of articles that were published with respect to each of the 42 categories across our dataset. From this, it can be seen clearly that "Politics" is the most common news category from 2012 to 2022, followed by "Wellness" and "Style and beauty" whereas "Arts" is the least common.

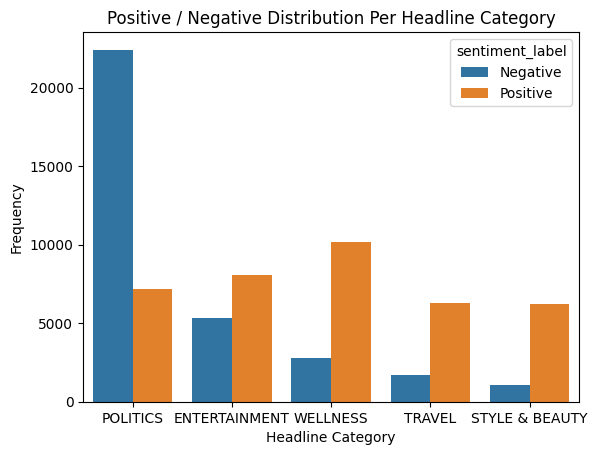
**Word Cloud :**



Above is the collection of the word clouds for each year from 2012 to 2022. The words with highest count includes "Trump","Donald", "President","Biden" which supports the fact that "politics" is the most frequently mentioned category.It could also relate to the "entertainment".

After "Politics" related words, we can see the prevalence of the words like "COVID", Coronavirus", "Vaccine" which are related "wellness"

**Sentimental Analysis Graph:**



**Clustering Plot using PCA:**

