# Contributors: Dylan Kakkanad, Mahika Bhartari, Rajashree Ramaprabu, Sahasra Konkala

# 

# Exploratory Data Analysis

As part of our initial EDA,

To analyze the news articles dataset, a dataset of approximately 2.7 million rows. From this extensive dataset, we employed the stratified random sampling method to derive a representative subset of approximately 27,000 rows ([1.0](#_qyjuqb88lvy)). Followed by analyzing the data structure of our data frame and then moved to check for null, duplicate, and invalid values. Duplicate and null values present in the article column are removed. We explored the number of articles published and the average count of words used by various publishers and depicted it using bar plots([1.1](#_1xaqpo7wam2o)). We have also visualized the number of articles available in distinct sections([1.3](#_imer3bh4vvw8)) published annually. We have created a word cloud model([1.5](#_px0z0zlb4lyu)) to showcase the most frequent words in the title, and article columns. In our further analysis and processing, we will observe how the recurring terms evolve in the word cloud.

# Pre-processing

Preprocessing mainly focussed on cleaning and preparing the ‘article’ column. The new processed column is called ‘processed\_article’([2.1](#_gmbpt8ku9bnq)). To reach this stage we first cleaned the articles by removing hyperlinks, numbers, punctuation, and non-English characters found during EDA. Since our goal is to cluster news articles based on the subject matter in the article, numbers (not in words) would be redundant or, worst case, bias the clusters and were subsequently dropped. We also used word tokenization and filtered for stemmed English words within the nltk English corpus excluding stopwords. The number of words was thus reduced to 20,000. Through the previous iteration found in phase-2 we were able to identify and remove words like ‘said’, ‘would’, ‘also’ etc.

# Analysis Plan

Our primary objective was to cluster news articles effectively and find distinct clusters. Given the lack of a well-defined metric, we adopted an iterative approach to achieve this goal. The first step in our methodology was to reduce the dimensionality of our dataset, which initially had over 20,000 features. We employed Principal Component Analysis (PCA) for this purpose. PCA allowed us to identify and retain the top 1,000 principal components that accounted for approximately 50% of the variance in our data ([3.1](#_7l5jyy3mbird)). This reduction was significant as it simplified our dataset while preserving its essential structure.

With a more manageable set of features, we then applied the KMeans clustering algorithm. The choice of KMeans was motivated by its efficiency and simplicity, making it a suitable choice for our initial exploration of the data. To determine the optimal number of clusters, we utilized the elbow method ([3.2](#_52thnwwace1m)) and silhouette plots ([3.3](#_x3256rt2ng7t)). Since we were not able to infer the exact number of clusters using the elbow method, we opted for a silhouette plot that provided us with the optimal number of clusters to be used. Both these techniques provided us with a quantitative means to assess the quality of our clustering.

We identified the top 30 words with the highest TF-IDF scores in each cluster. This step was crucial as it provided us with a qualitative understanding of the content of our clusters and is the only metric to assess the clusters. We also used cosine similarity to find top 5 similar documents to a given output. Similarly, we have also identified the top 30 words clusters with different n-gram ranges to cross-validate how the clustering model performs for different approaches. To reduce complexity in hyperparameter tuning for higher n-gram ranges we utilized subsetting based on year and publishers. These data subsets help not only with complexity management but also highlight business use cases better.

We employed the Non-negative Matrix Factorization(NMF) model to analyze article text, utilizing vectors derived from the TFIDF model. Our approach involved iterative refinement to determine the optimal number of topics. Initially, we conducted a comprehensive analysis of the entire dataset, extracting the top 10 topics along with their representative words. Subsequently, we conducted a year-wise examination of articles spanning from 2016 to 2020. This method allows us to track the evolution of trends over time and also helps in strengthening the model.

Utilized the VADER sentiment analysis tool from the NLTK library to analyze the sentiment of articles within finance-related clusters (identified as clusters 5, 6, and 9). By first filtering the original dataset to include only articles belonging to these finance clusters, we apply sentiment analysis to each processed article using VADER's polarity scoring mechanism. The sentiment of each article is evaluated and the sentiment scores obtained are then categorized into 'positive', 'negative', or 'neutral' labels. With this approach, we tried to understand the sentiment expressed in finance-related articles

In conclusion, our iterative approach, underpinned by strategic choices of PCA for dimensionality reduction, KMeans for clustering, and the elbow method and silhouette plots for determining the optimal number of clusters, allowed us to effectively cluster news articles. This methodology not only achieved our goal but also laid a robust foundation for future explorations in the field of text analysis, opening avenues for further refinement and optimization.

# Results

The clarity of themes within each cluster was our evaluation metric. After learning from our previous attempt we now tried it with a better feature set and increased the number of clusters to 15. Three clusters could not be placed into any groups by looking at the top words in the clusters. The remaining clusters, however, showed distinct themes like crime news, federal government news, national politics, business finance, financial markets, sports news, and the stock market([4.1](#_b7rkku6jtv39)). Our evaluation metric was the clarity of the themes within each cluster. out of 15 clusters, 12 had distinct themes with little overlap. However, the clarity of themes is also affected by sub-par clustering. Our clusters included at least 3 finance-related clusters which could all be combined into one cluster.

Adjusting the n\_grams range in TF-IDF vectorization resulted in similar clusters. Removing ineffective words(‘said’, ‘also’ etc) helped find new clusters like healthcare. Despite some unclear clusters, our initial findings are promising. We have performed unigram with range(1,2). The cluster order wasn't the same as TF-IDF, however, it yielded pretty similar results([4.2](#_sbyp28cs4rbm)) with few repetitive clusters contributing to finance and politics. Two of the resultant clusters are not clearly distinguishable to figure out the genre of it. Most of the clusters provided promising results for initial clustering with minimal overlaps. Moreover, we performed bigrams for the newspaper articles belonging to the year 2019, we saw articles related to politics in the first few clusters which were also overlapping. A really good cluster(=4) that was formed was articles that had “New York” ([4.3](#_7ec808i1jp4q))in it. We were able to see new clusters associated with social media names in it. Next, we also performed trigrams subsetting publication as “People”. Though there were some spillovers, the model did a satisfactory job at clustering different topics and we were able to find new clusters in entertainment and law and crime using the tri-gram([4.4](#_8ojli7wwvgue)).

Through the NMF model for all the articles, a wide range of topics including Financial News, Politics and Government, Lifestyle, Law and Governance, Company Financial News, Corporate News and Acquisitions, Global Trade and Economy, Stock Market, Legal Discussions, and Sports were obtained. For each year, we identified the top 10 topics([4.5](#_7cr71u48h1x9)), revealing significant overlap across years in key areas such as Politics and Government, Stock Market, Financial News, Global Trade and Economy, and Health Care. Notably, each year introduced at least one new topic reflecting the most trending theme, such as Trump’s campaigns in 2016([4.6](#_8iojrteg2ylz)), Russian interference in 2017([4.7](#_wc8ttuduuw21)), Trade wars in 2018([4.8](#_9dqdf71t1o6o)), the Oil crisis in 2019([4.9](#_crl872a6uhqc)), and the covid-19 pandemic in 2020([4.10](#_ki8bk9xsy15s)). This adaptive approach ensures our model remains relevant and accurate, capturing current trends and reader interests. News channels can also utilize these insights to update their trending section.

Our sentiment analysis on finance-related articles revealed a wide range of sentiments. Despite market volatility, many articles showed neutral sentiment, indicating balanced reporting. Peaks in positive sentiment could correspond to favorable market developments, while dips might align with negative events. These insights can help businesses gauge market sentiment and inform strategies, forecasting, and product development in the finance sector.

# Limitations

# 

* POS Tagging and NER: The potential of POS-tagged and weighted TF-IDF models remains untapped. Future work could explore creating multiple models based on different weights, potentially enhancing results.
* Hierarchical Clustering: The high time complexity (O(n³)) and space complexity (O(n²)) of hierarchical clustering posed challenges with our dataset of 26,000 rows. Future efforts could explore efficient sampling methods or alternative algorithms for large datasets.
* Word2Vec: While Word2Vec’s extensive list of 3 million words and phrases is a strength, the 300-dimensional feature space is resource-intensive. Future work could investigate dimensionality reduction techniques to manage this challenge.
* Sentiment Analysis: A challenge with sentiment analysis across clusters belonging to unique sectors is that each genre may have its unique keywords, vocabularies, and contextual nuances, in such cases it might potentially lead to less accurate sentiment classification.

# Coding Contribution

For contributions, a detailed table is provided in the appendix([fig 0](#_o73j66rpjnee)). A **kanban board** used for the project can be found at this [link](https://climbing-savory-75f.notion.site/52925d7515424d7f952b5da126212fca?v=4b28f448a65d44b594714ffc9bdf6fbe). Everyone equally contributed to the analysis and slides.

# GitHub Project [News Article Text Analysis](https://github.com/mahikabhartari/BA820-Project-News-Article-Text-Analysis)

# Appendix

#### Fig 0: Coding Contribution

| Name | Members | Contribution per member |
| --- | --- | --- |
| EDA | Mahika, Rajashree | 50%,50% |
| Pre-processing (incl stem) | Dylan, Sahasra | 50%,50% |
| Bag of Words | Dylan, Sahasra | 50%,50% |
| TF-IDF model + clustering | Dylan | 100% |
| TF-IDF model hyperparam | Mahika, Rajashree | 50% 50% |
| Cosine similarity | Dylan | 100% |
| Ngram (for range in 1.2) + clustering | Mahika, Rajashree | 50%, 50% |
| Ngram (for range in 2,3) + clustering year wise | Rajashree | 100% |
| Ngram (trigrams) + clustering publication wise | Rajashree | 100% |
| NMF model(incl sub-models) | Sahasra | 100% |
| NMF all visualizations | Mahika | 100% |
| Sentimental Analysis | Rajashree | 100% |
| Consolidation | Dylan, Mahika, Rajashree, Sahasra | 25%, 25%, 25%,25% |
| PCA scatter plots | Mahika | 100% |

# 

#### Fig 1.0: Data Dictionary

| Name | Description | Data Type |
| --- | --- | --- |
| date | Date of article publication | str |
| year | Year of article publication | int |
| month | Month of article publication | float |
| day | Day of article publication | int |
| author | Article author | str |
| title | Article title | str |
| section | Section of the publication in which the article appeared | str |
| article | Article text | str |
| url | Article URL | str |
| publication | Name of the article publication | str |

#### Fig 1.1: Number of Articles by Publication

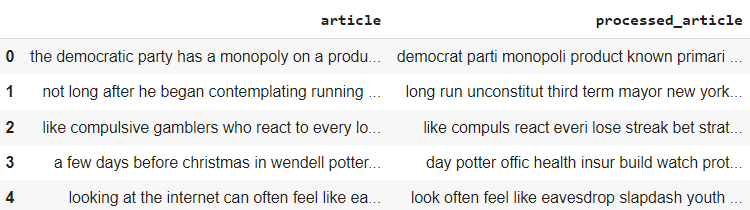
#### 

#### Fig 1.3: Distribution of Articles Across Sections

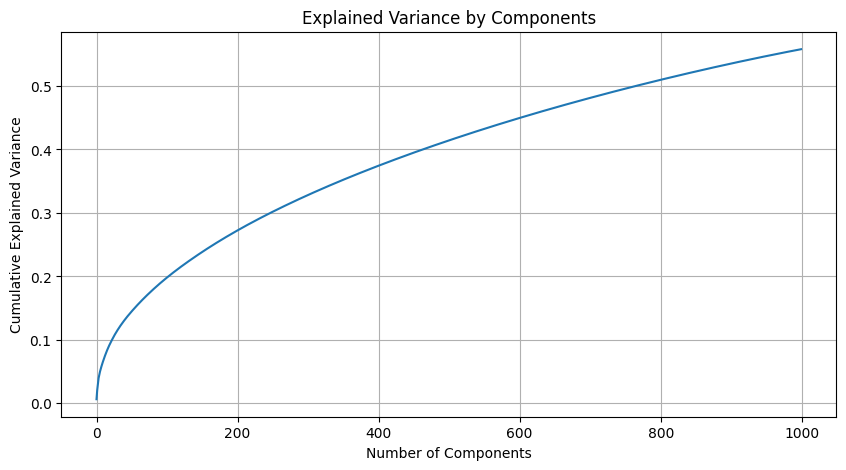
#### 

#### Fig 1.5: Word Cloud of the column ‘title’

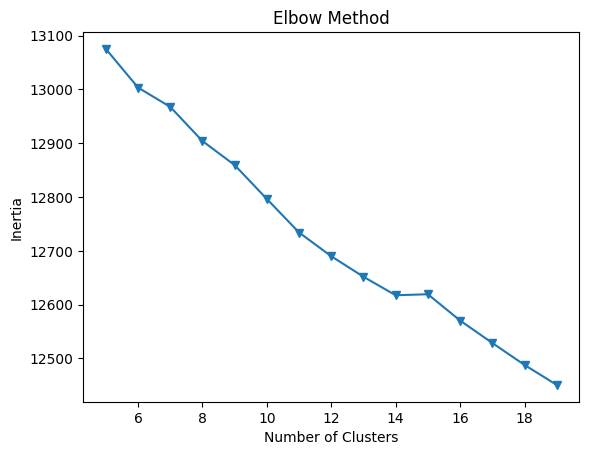
#### Fig 2.1: article column pre and post processing



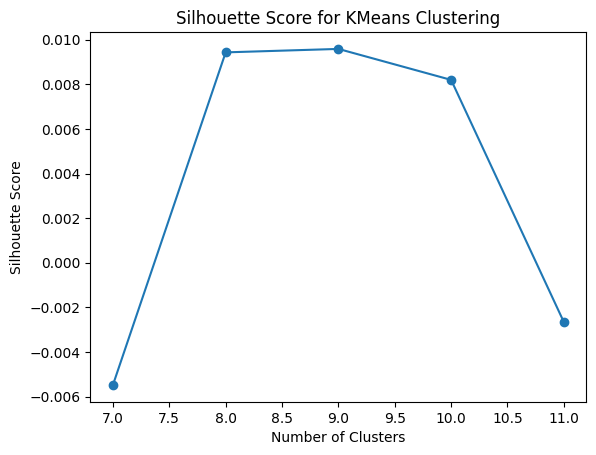
#### Fig 3.1 PCA:



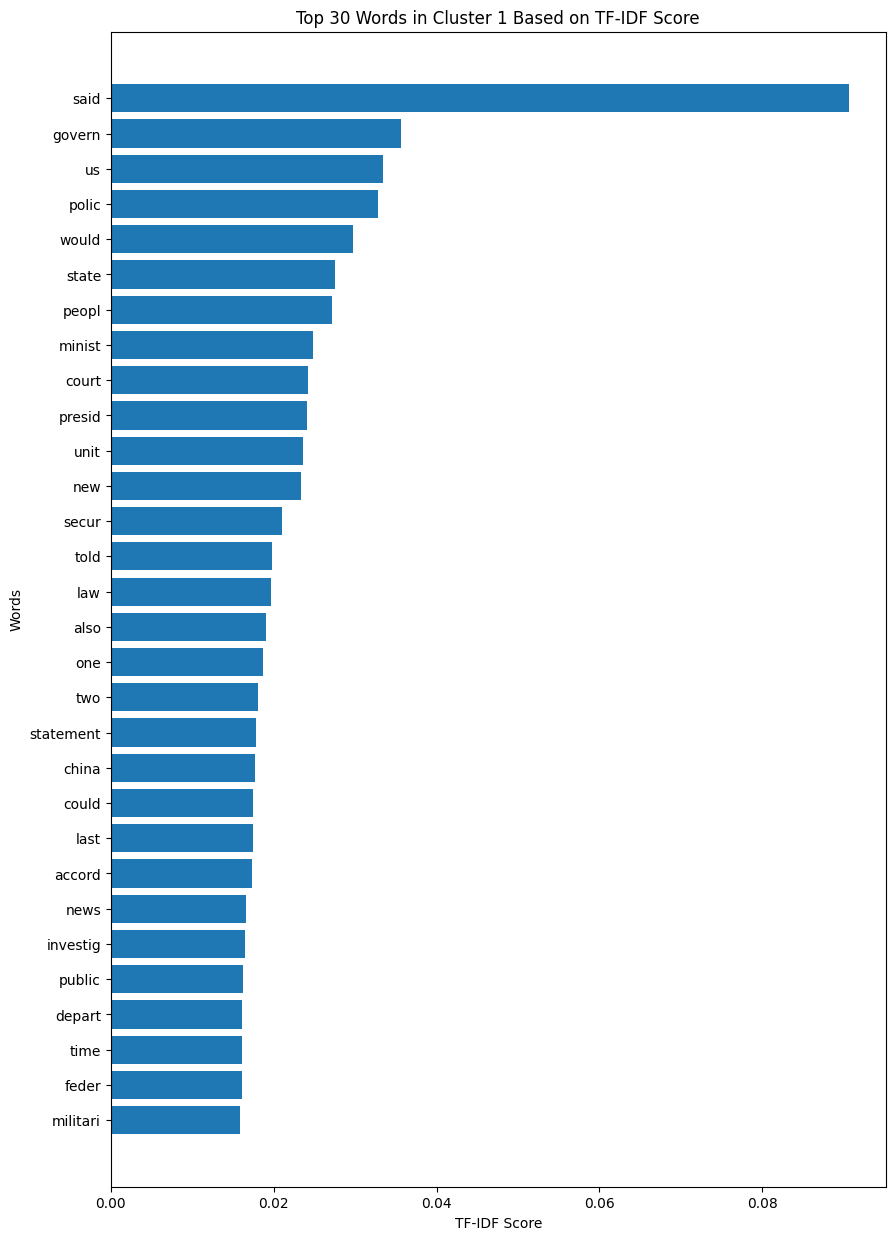
#### Fig 3.2 Elbow Method

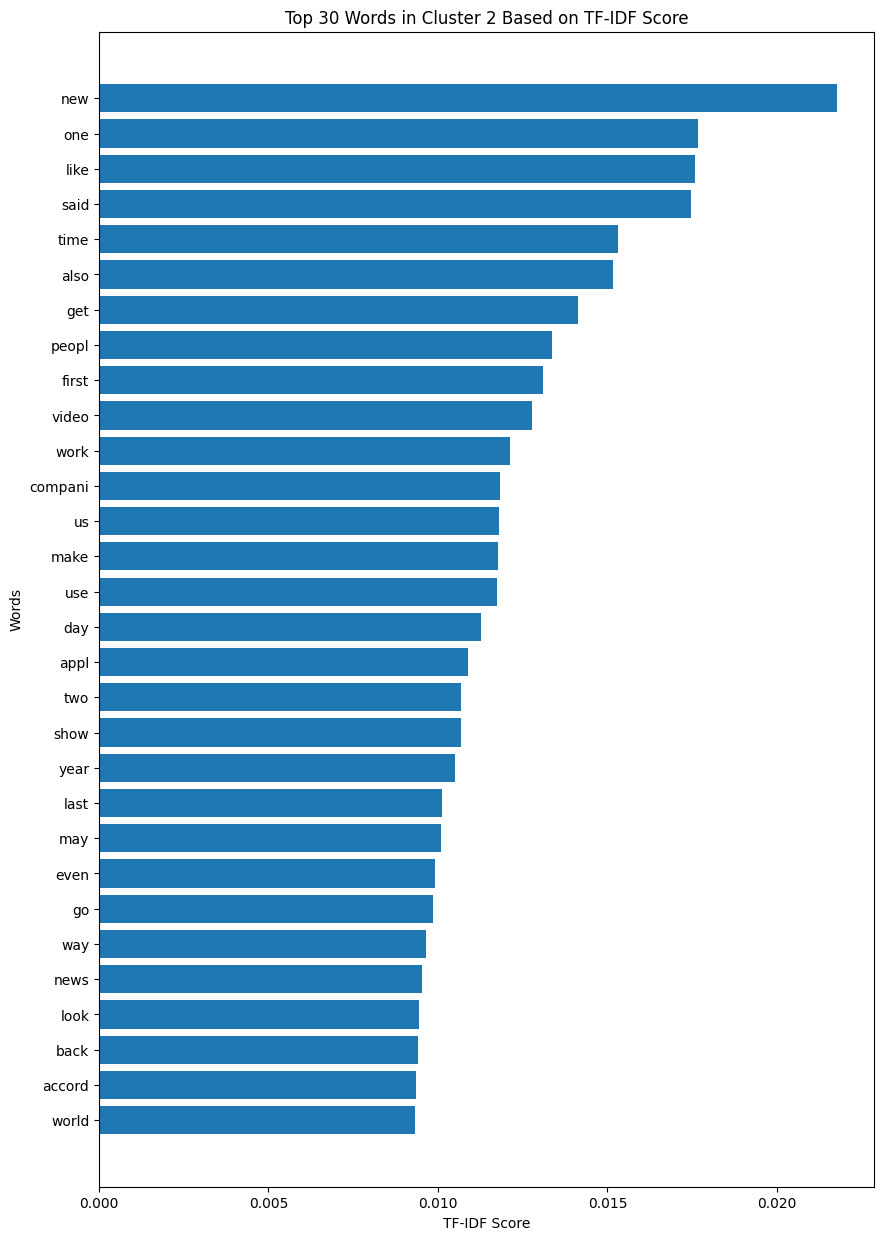
******

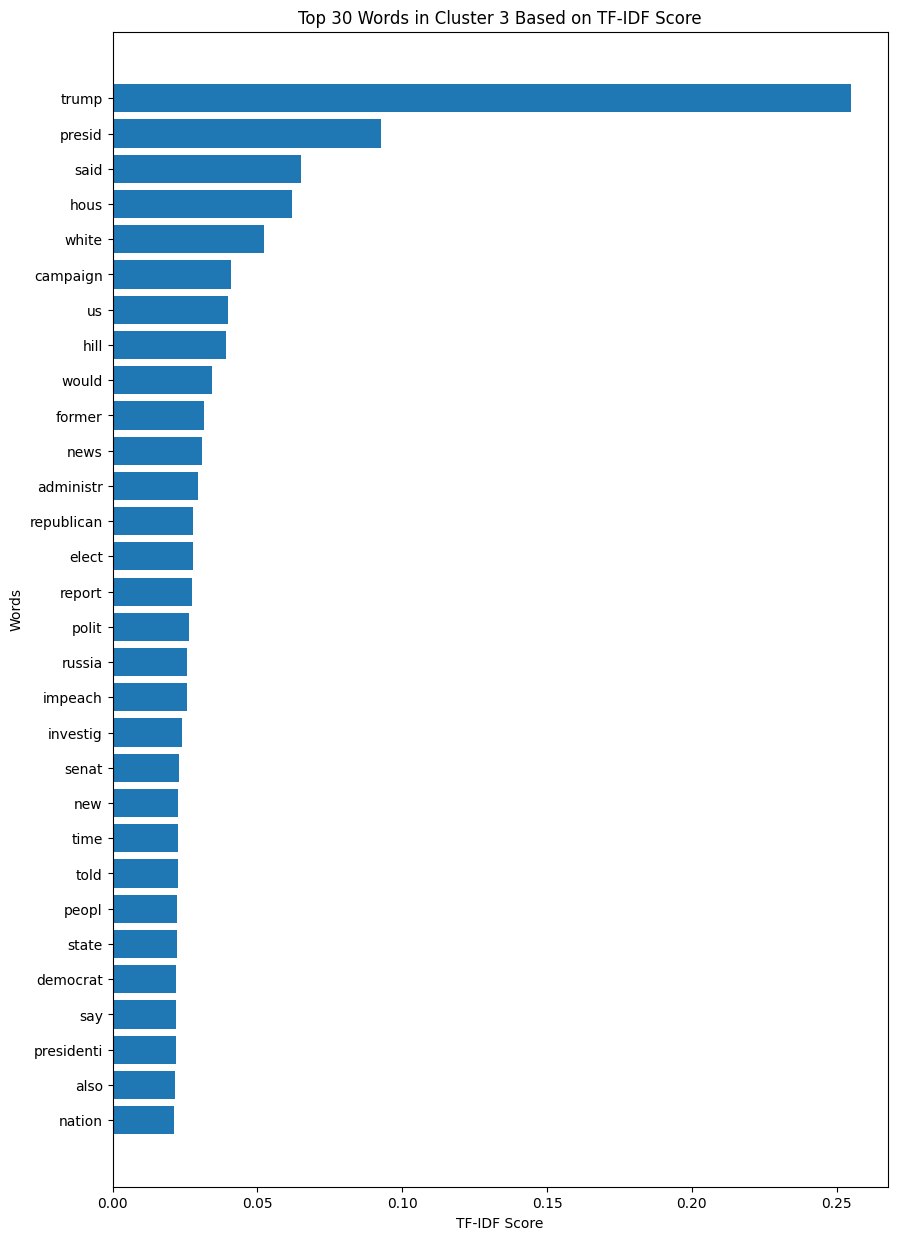
#### ***Fig 3.3*** Silhouette Plot:

******

#### Fig 4.1 Clustering results through KMeans



******

******

#### 

#### 

#### 

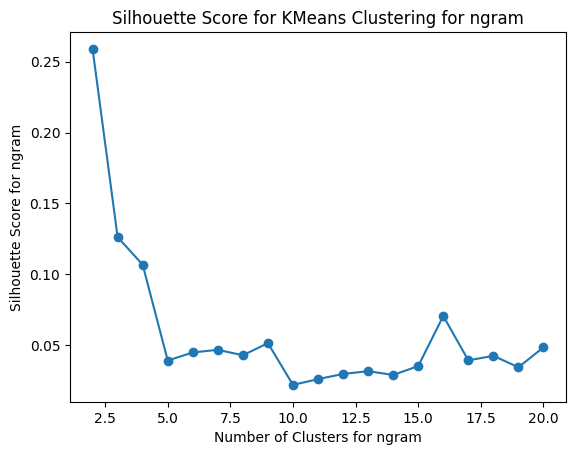
#### 

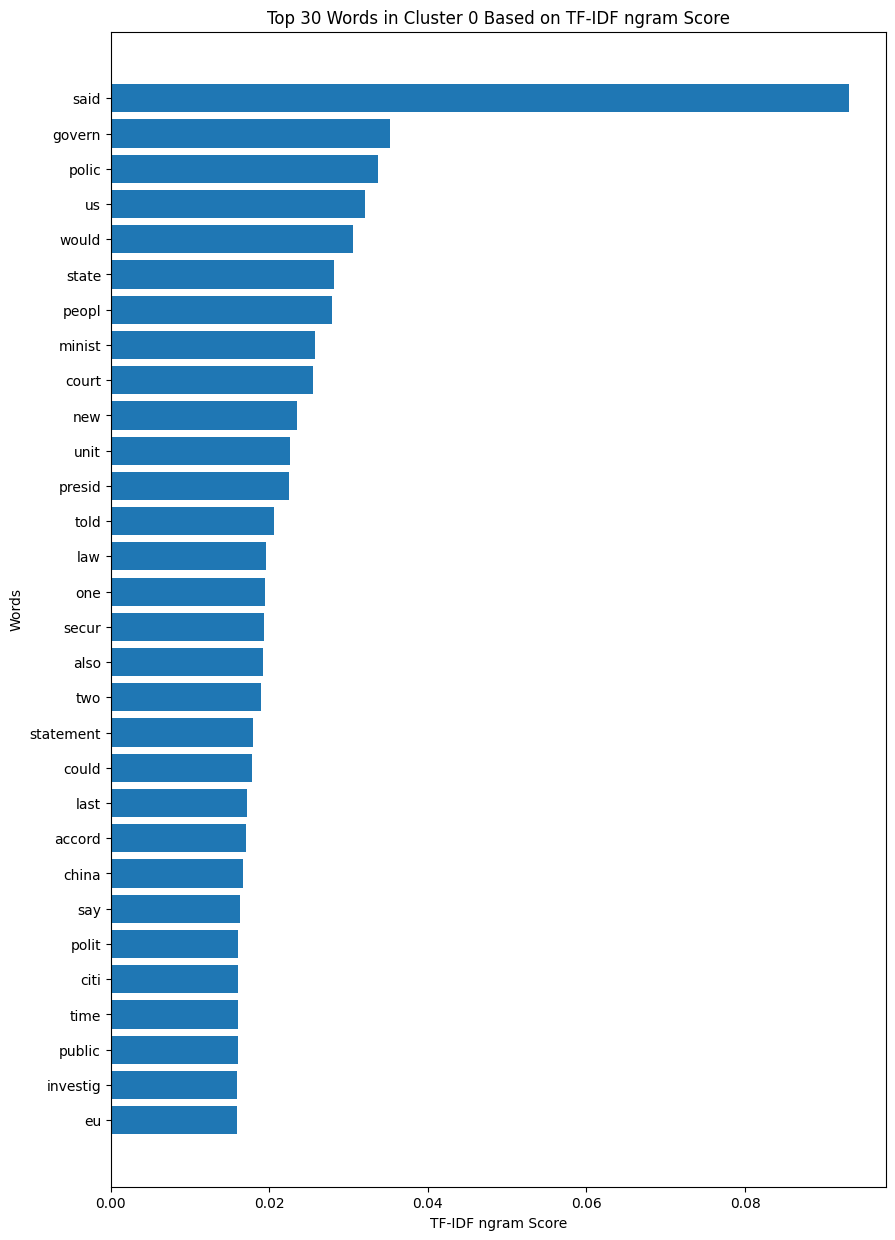
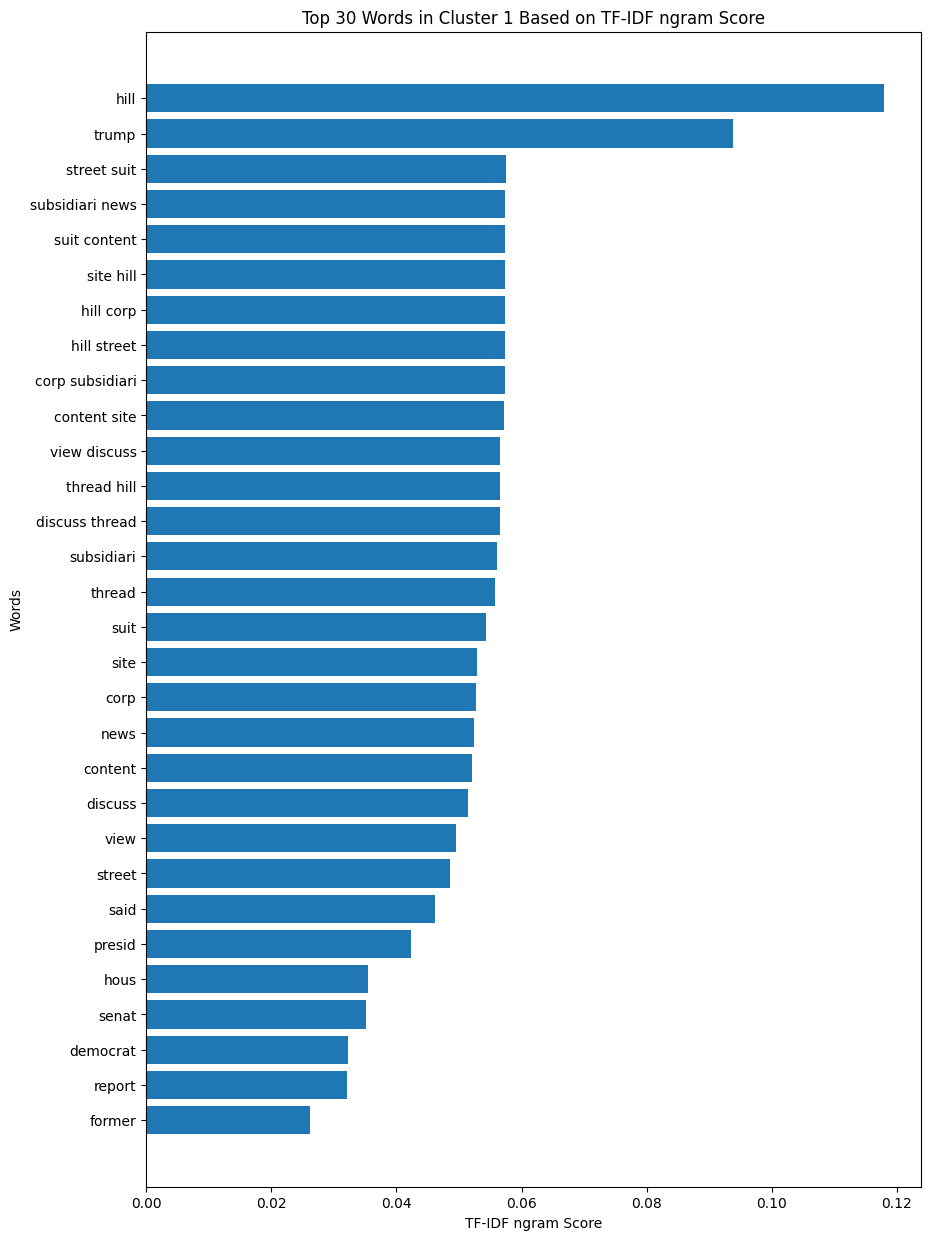
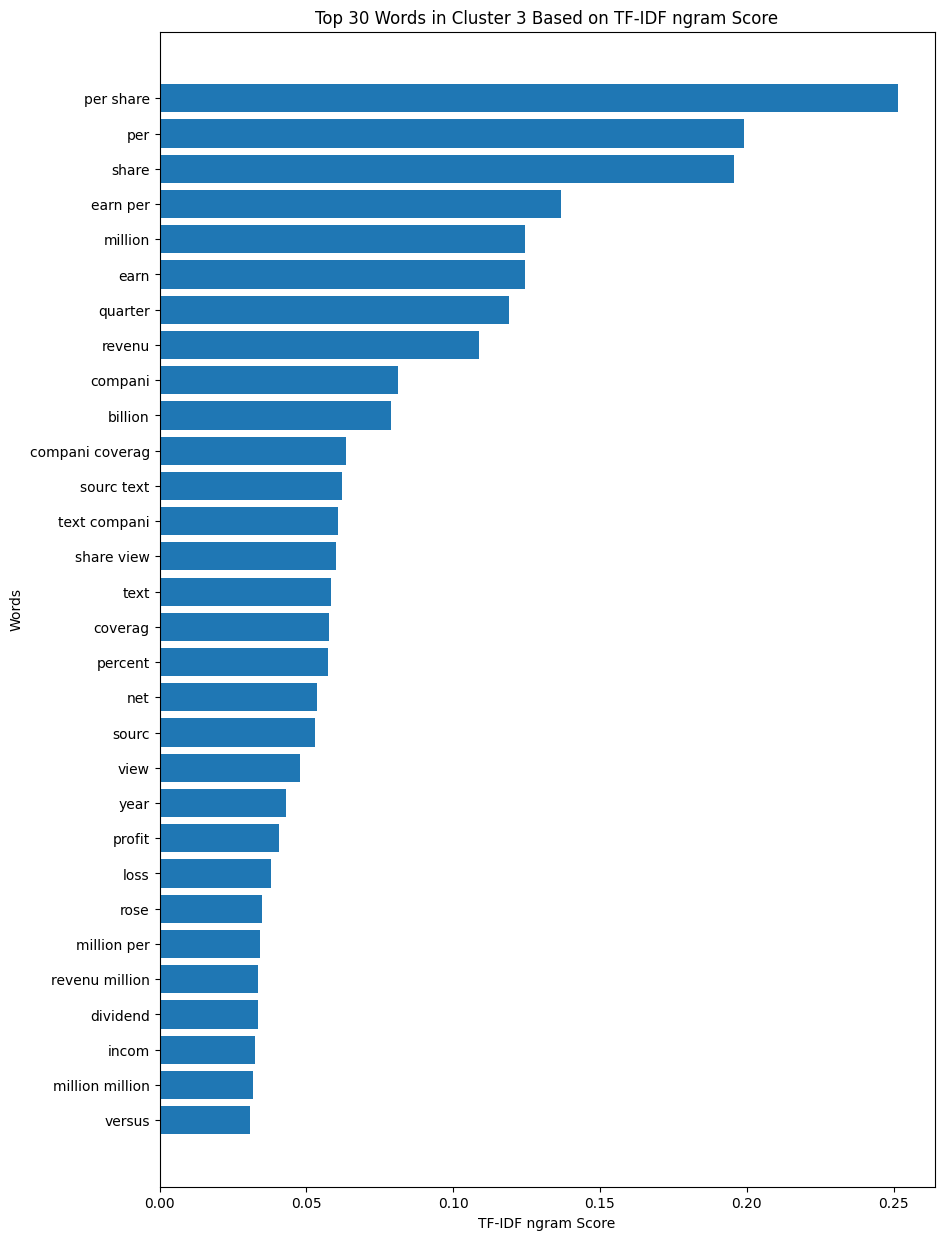
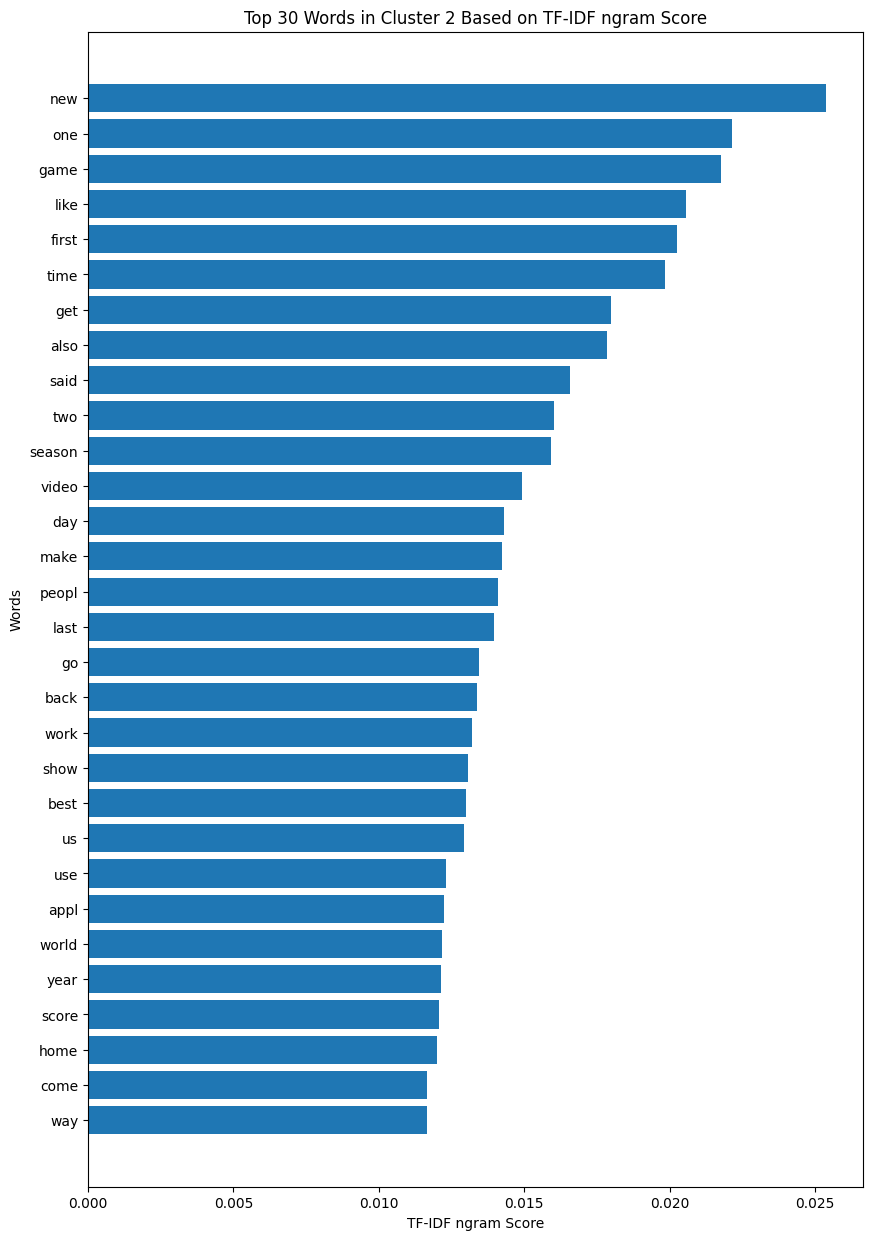
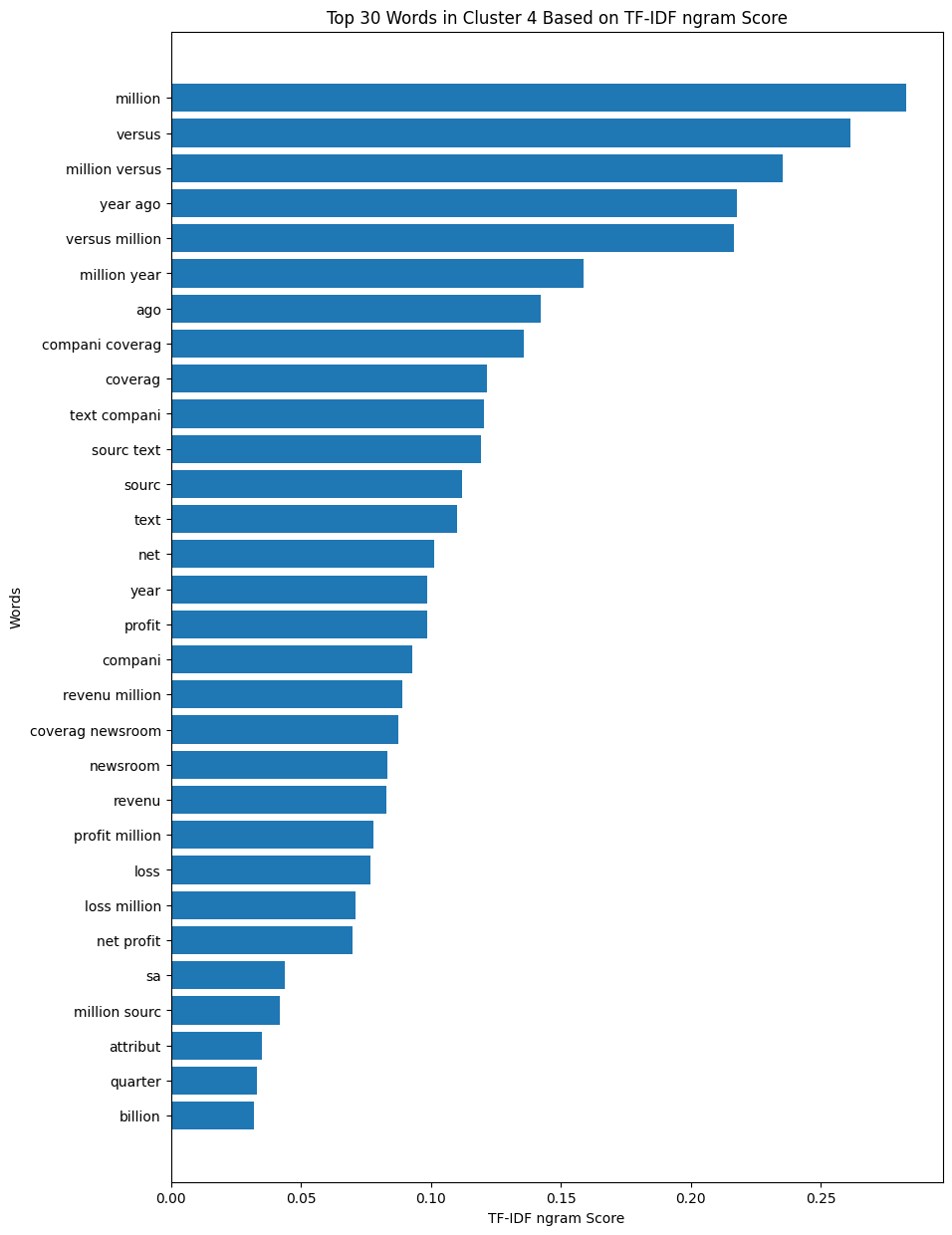
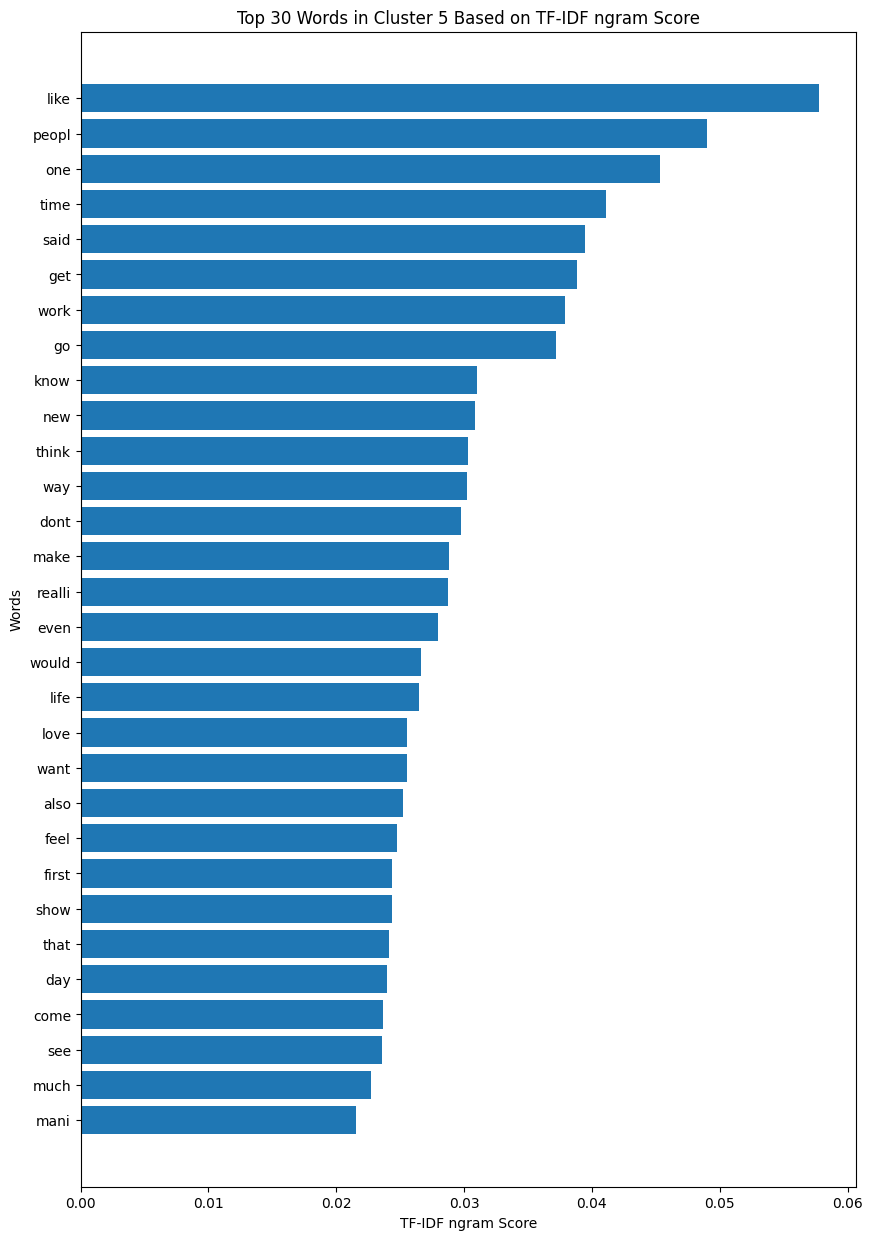
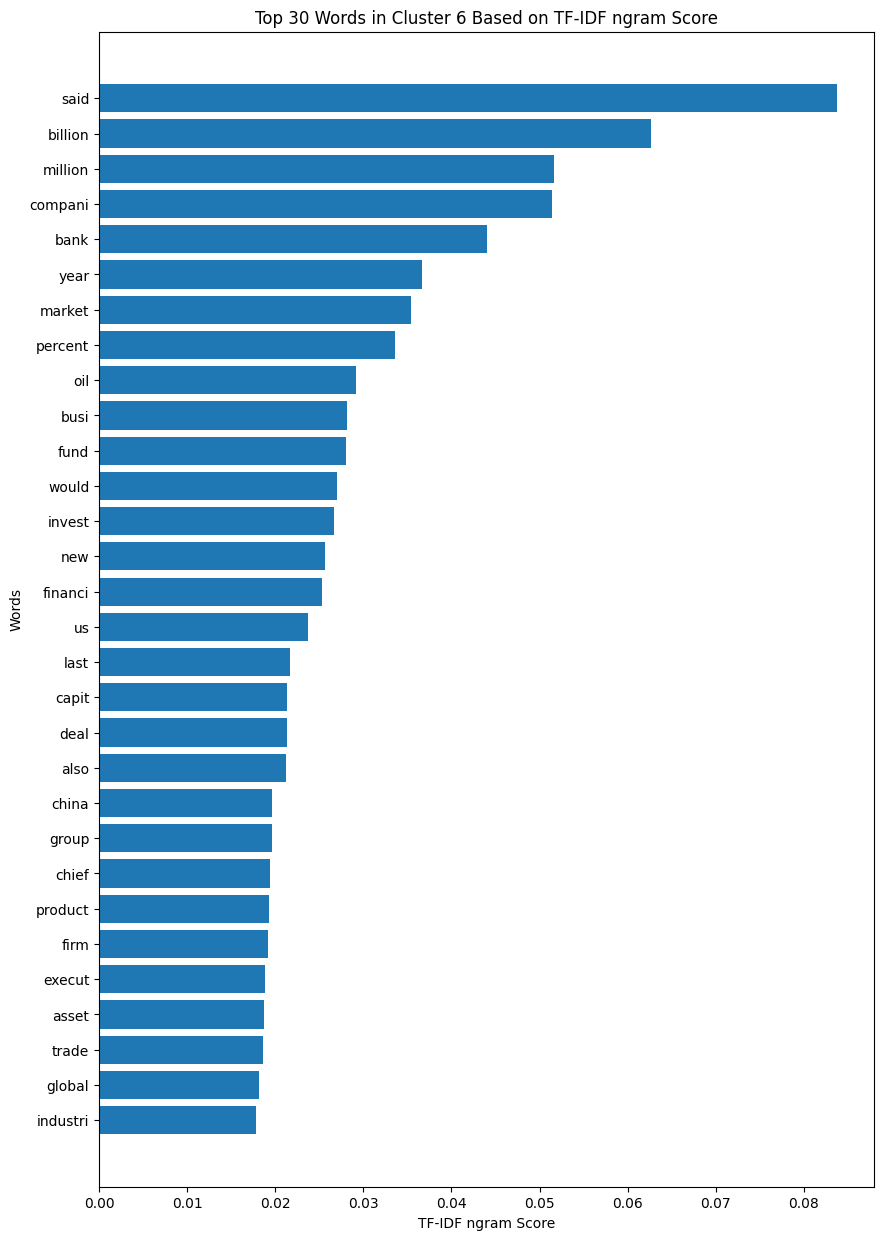
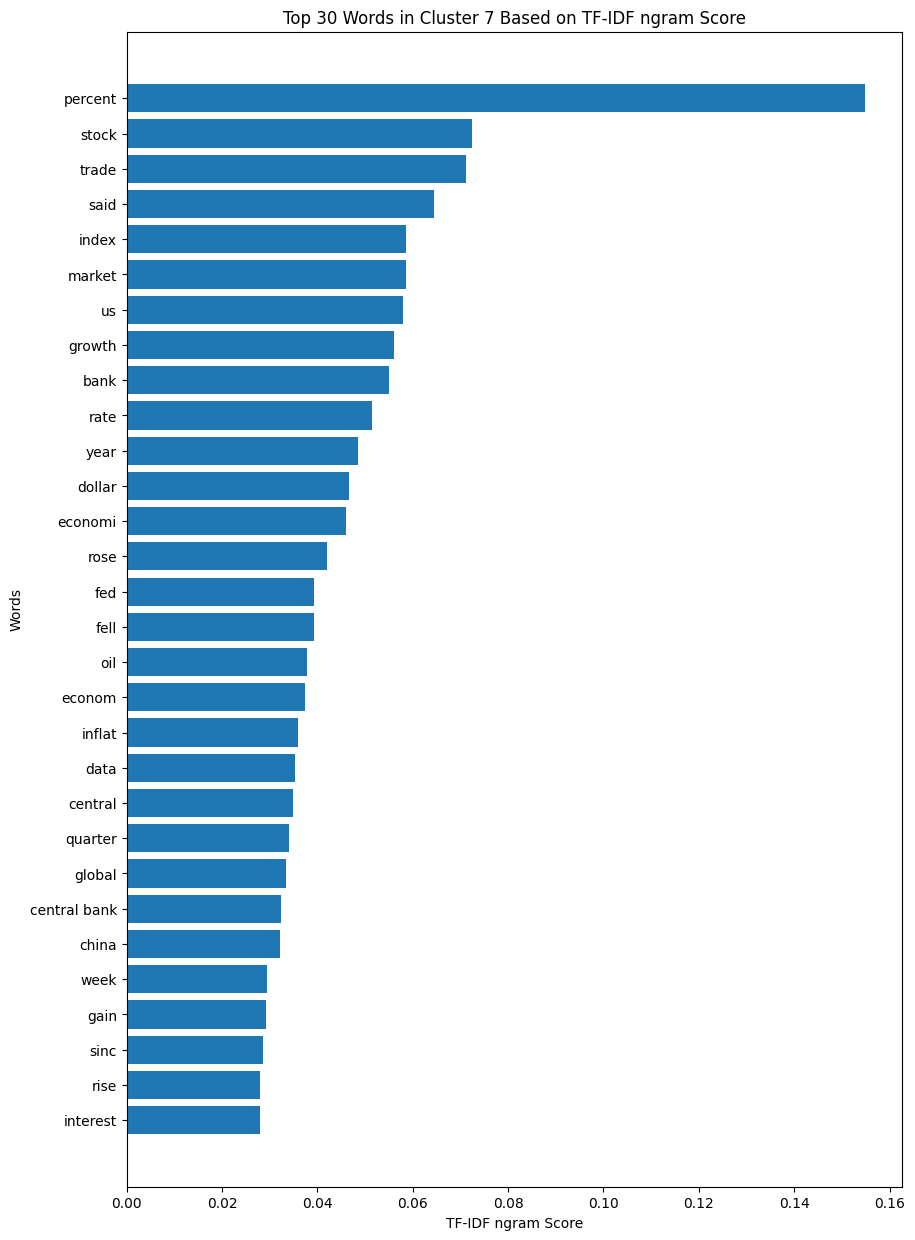
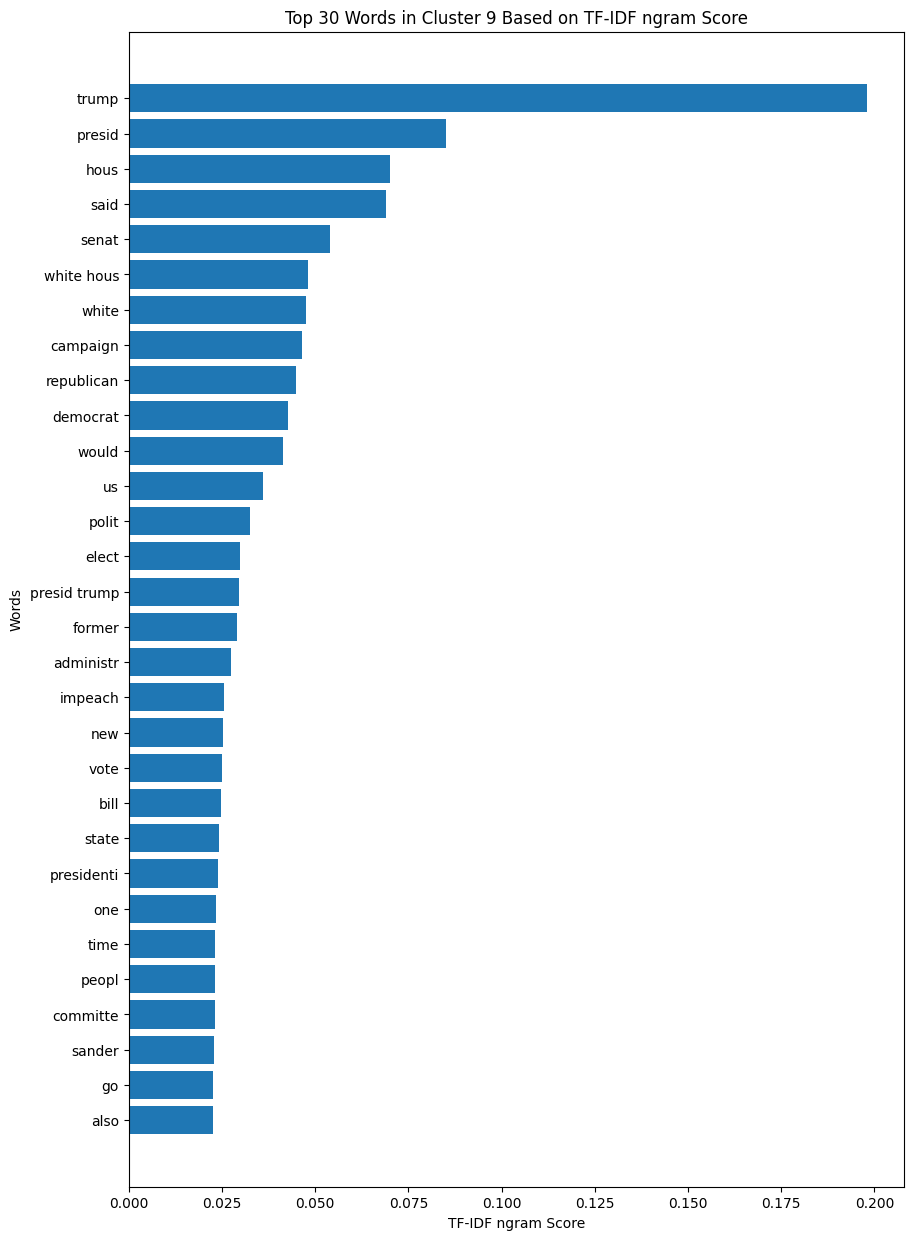
#### 

#### Fig 4.2 PCA Elbow Method, Silhouette Plot (ngram)

#### 

#### 

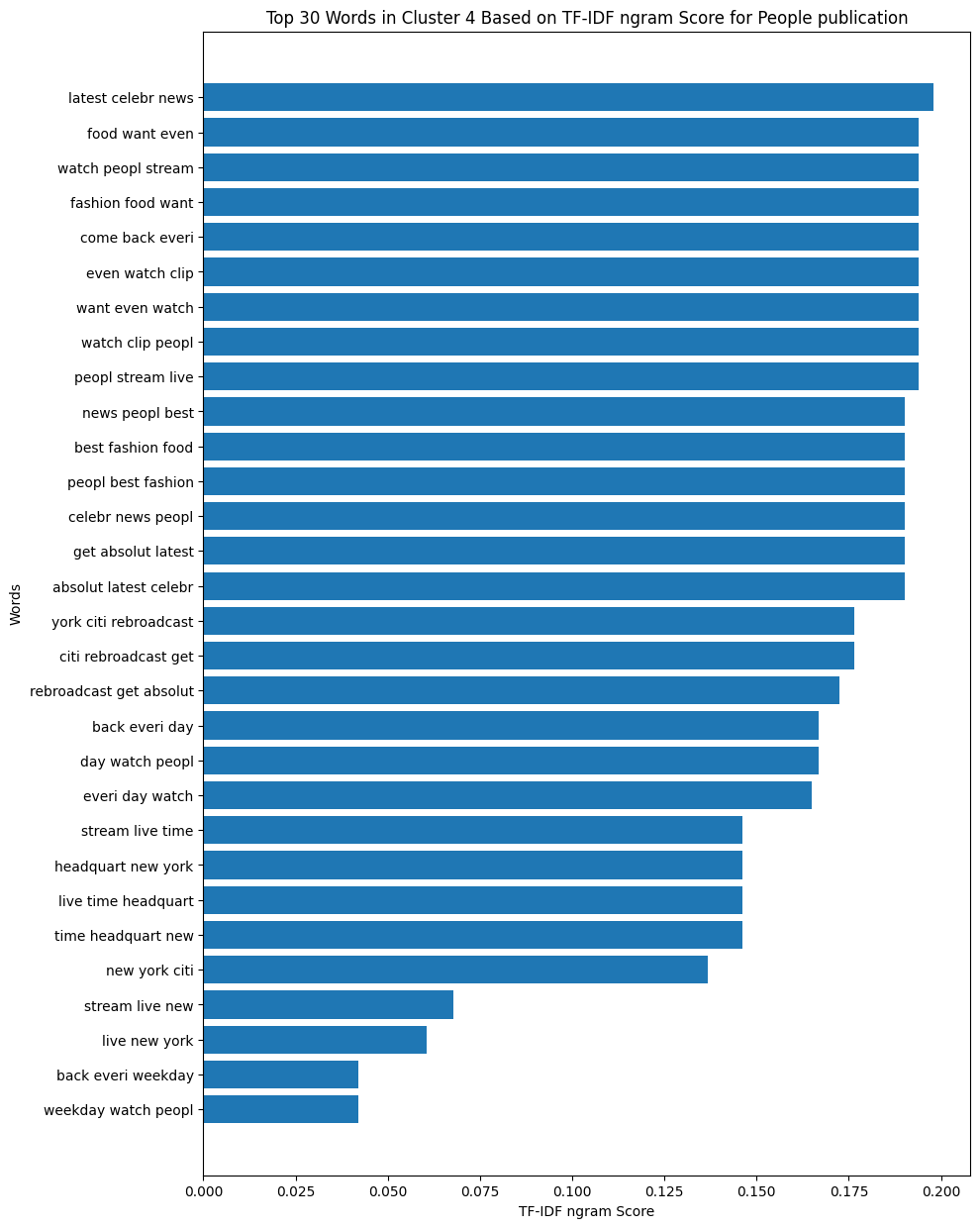
****

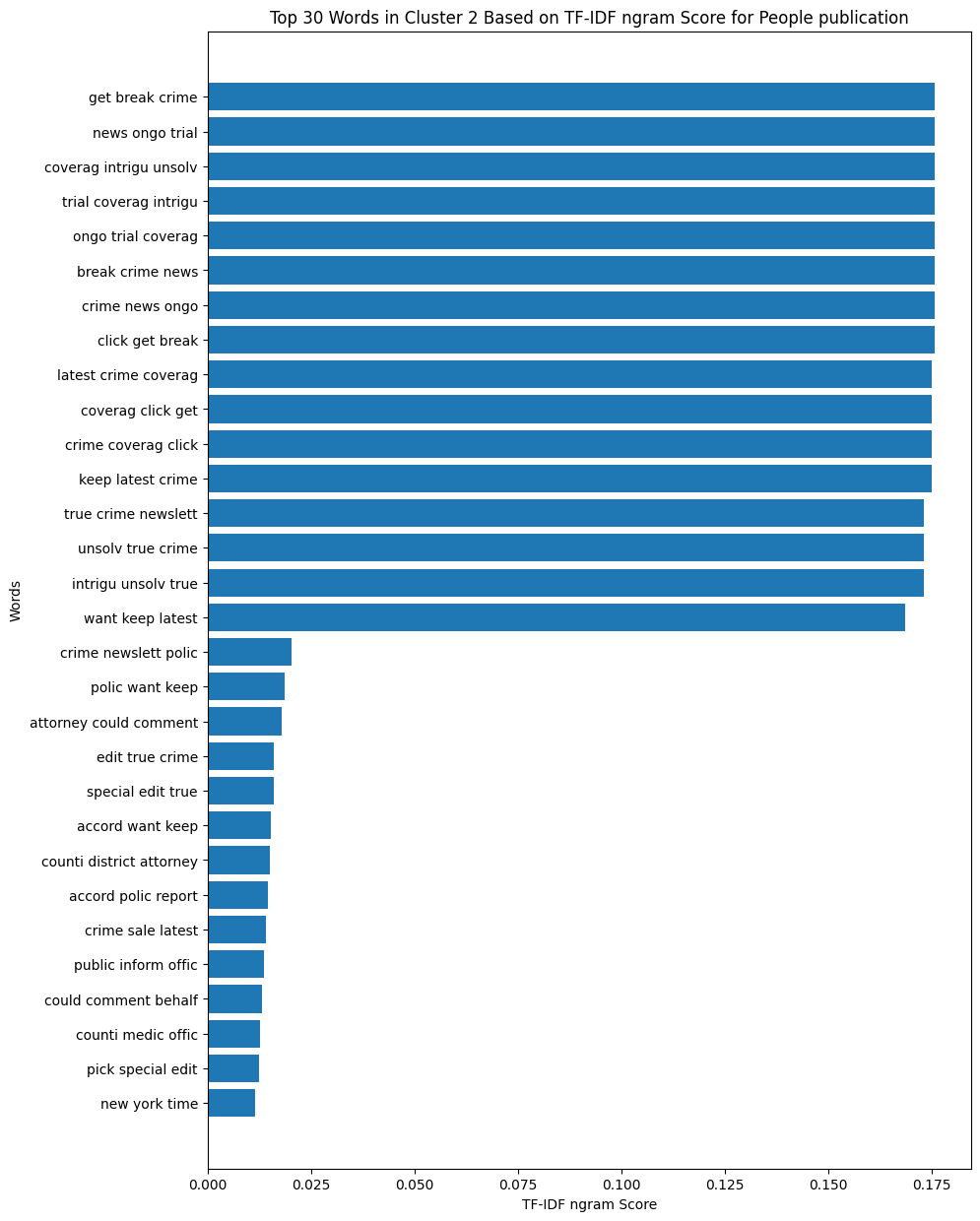
****

#### Fig 4.3 NewYork cluster by sampling 2019 data with (2,3) variate range

#### 

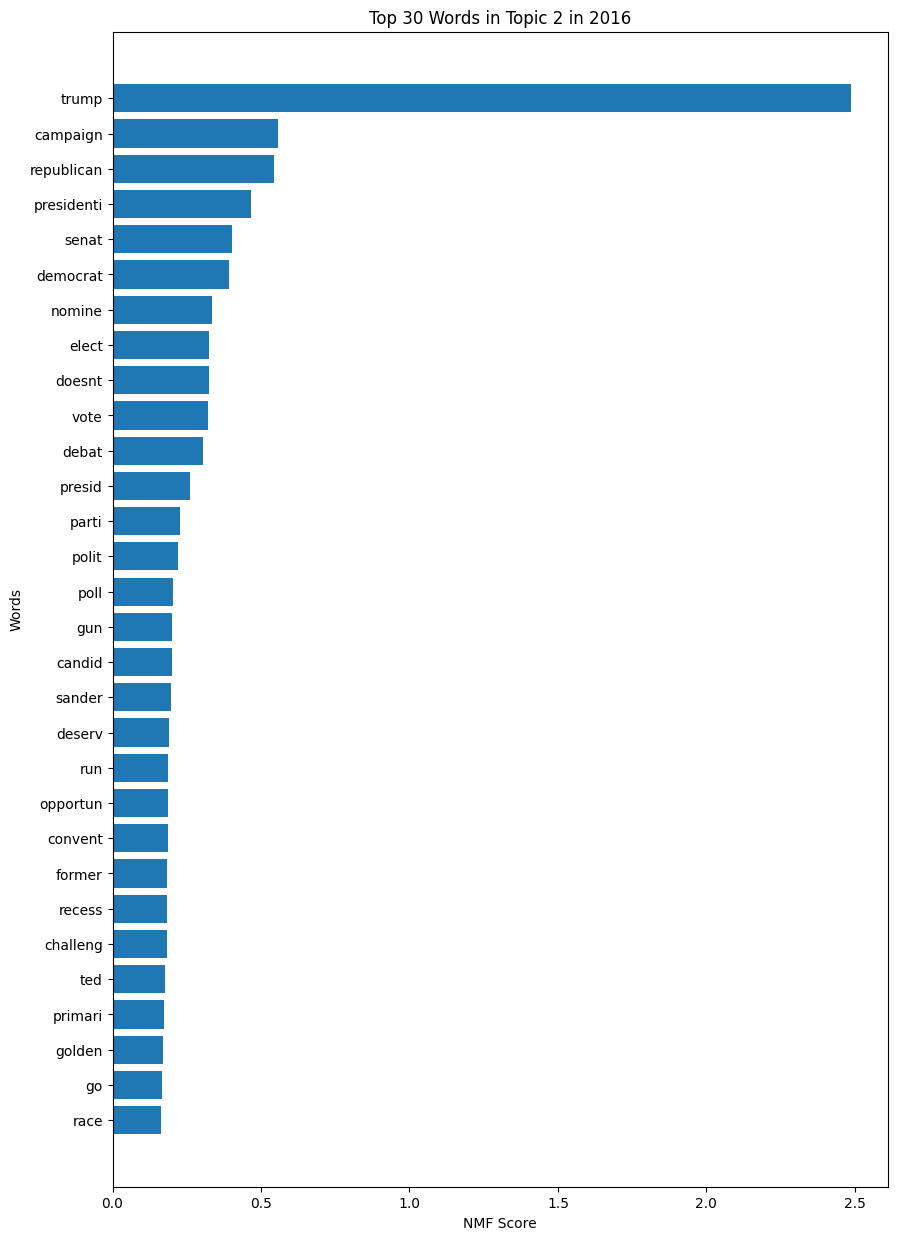
#### Fig 4.4 Entertainment and Law and Crime cluster using tri-gram

****

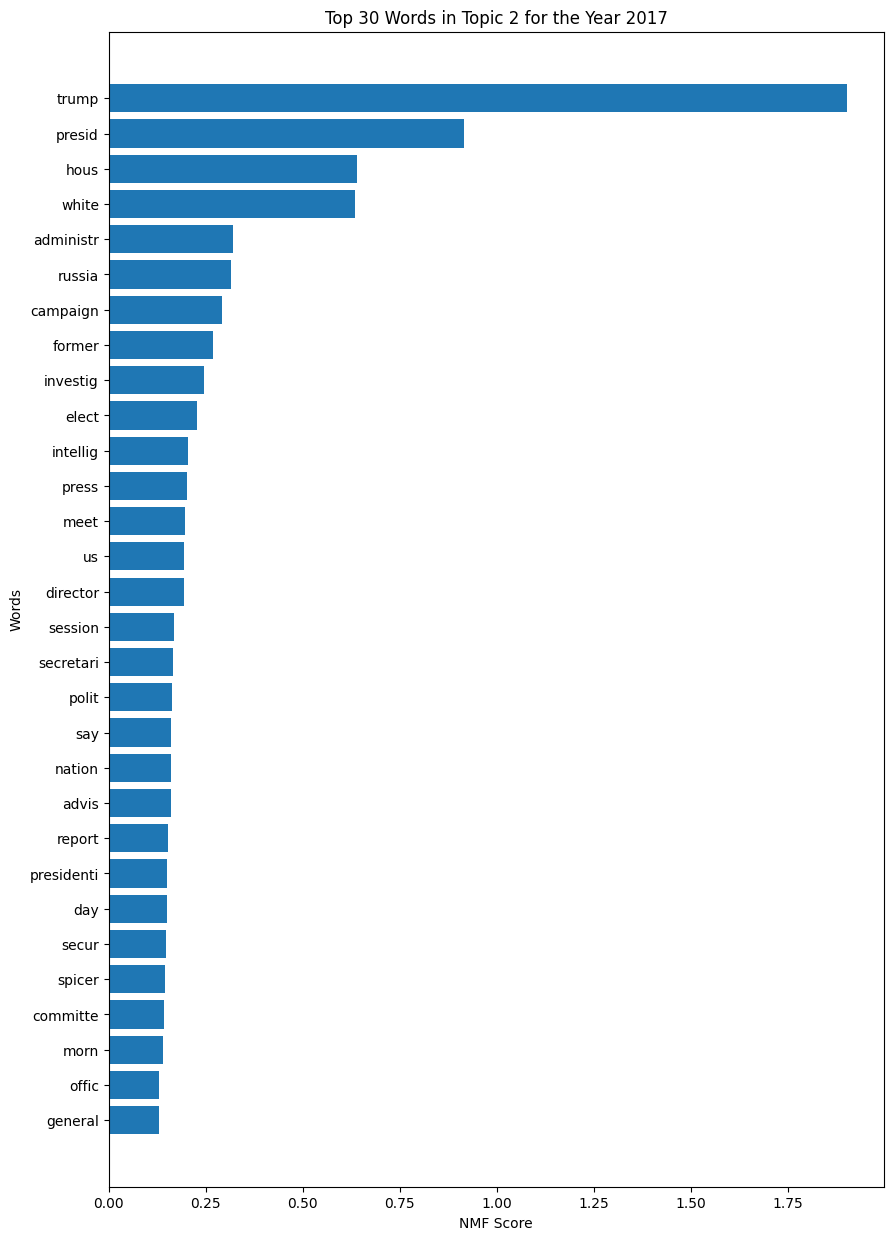
****

#### Fig 4.5 *NMF topic modelling:*

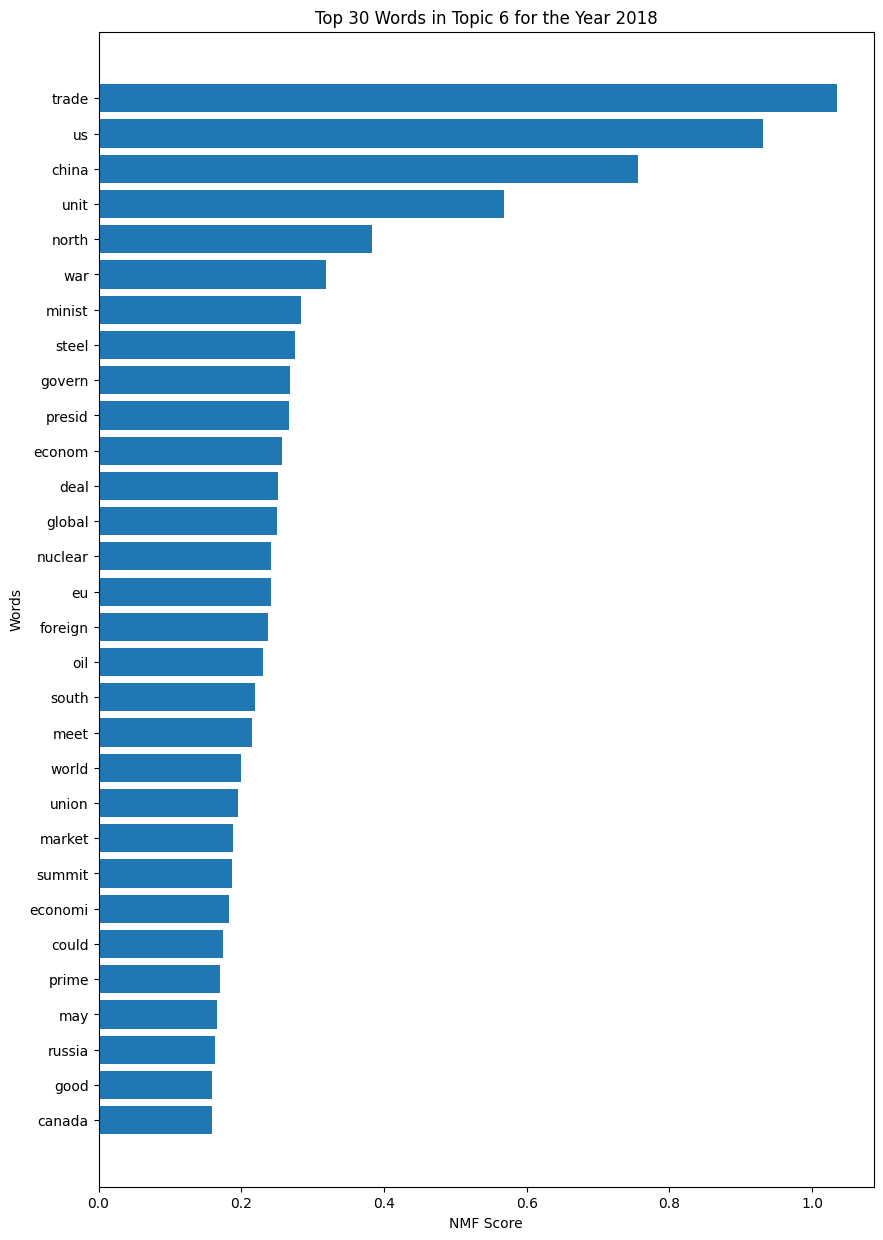
#### Fig 4.6 NMF topic modelling-New topic obtained from 2016:

****

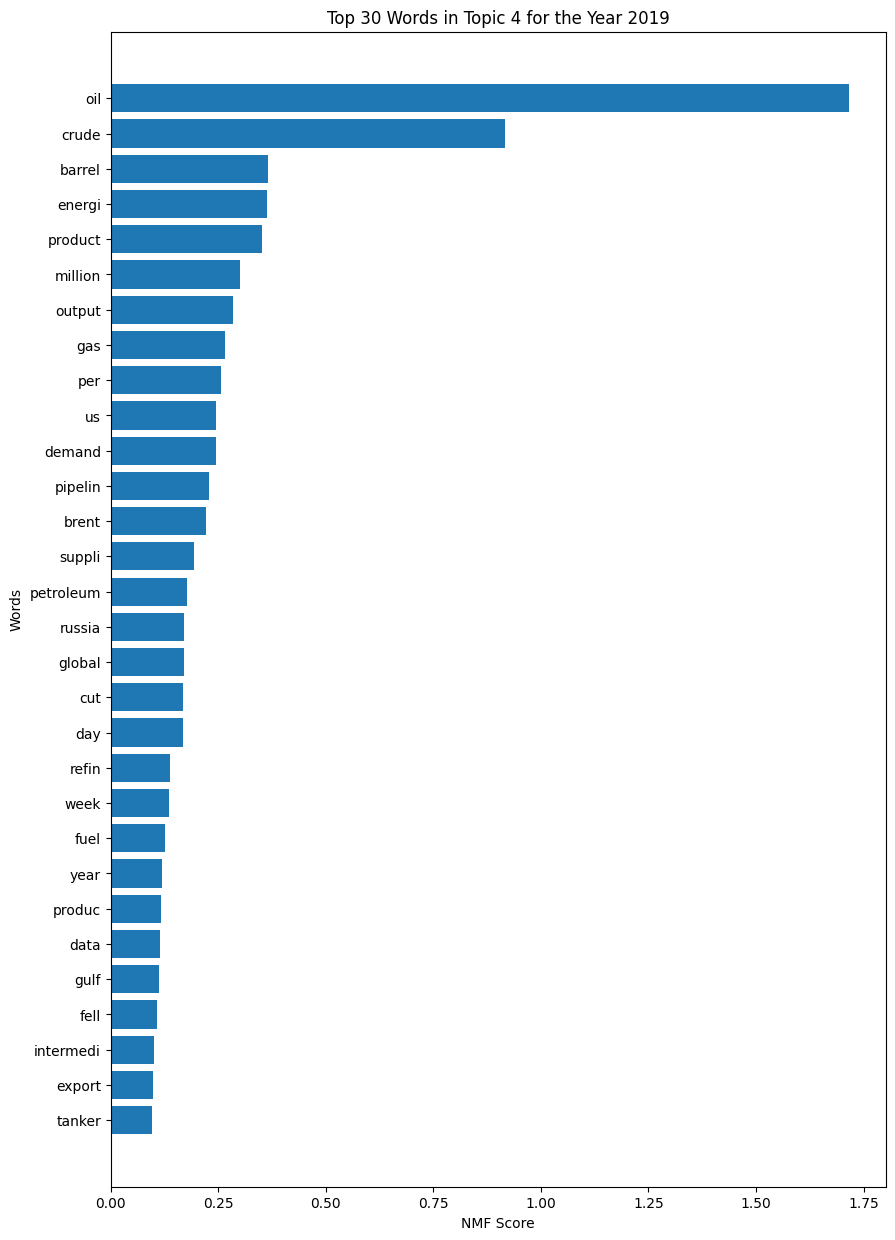
#### Fig 4.7 NMF topic modelling-New topic obtained from 2017:

****

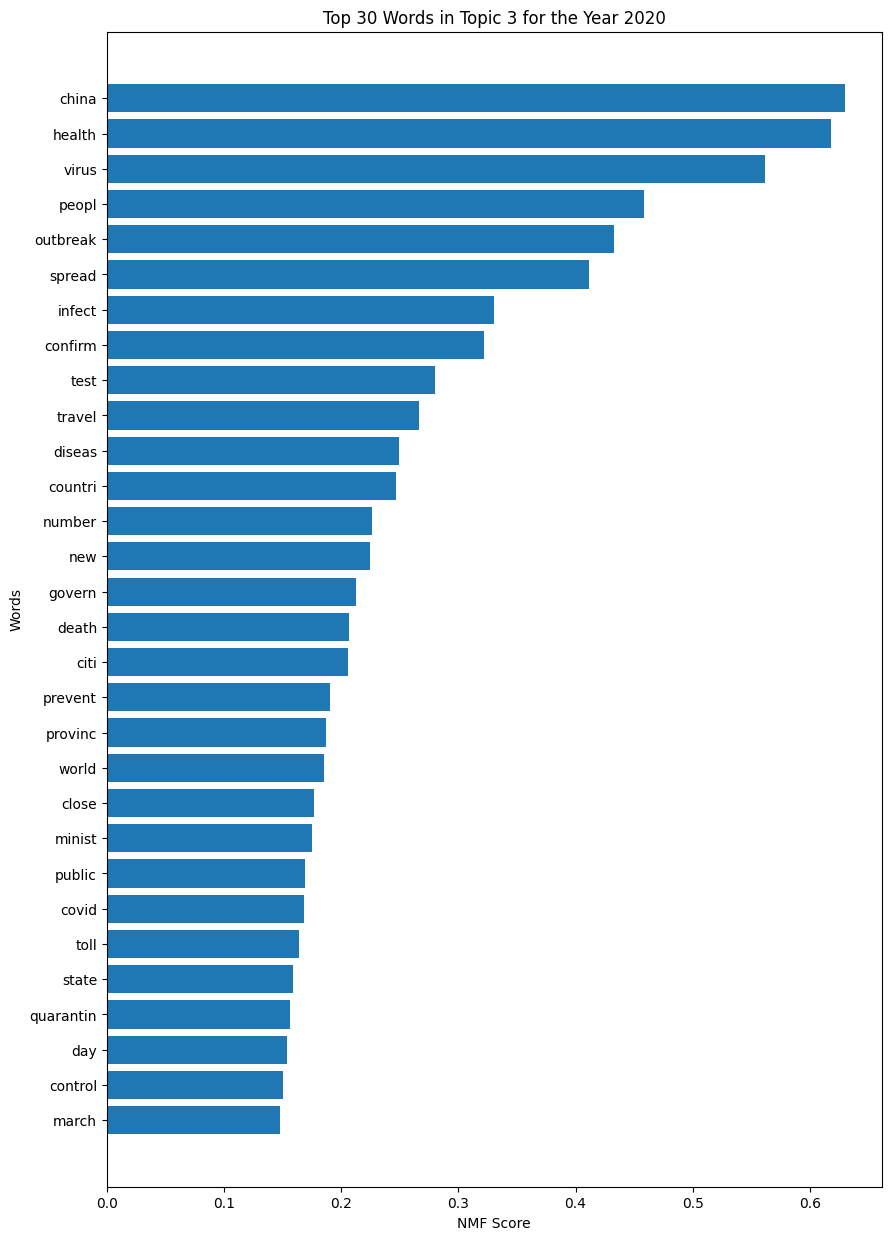
#### Fig 4.8 NMF topic modelling-New topic obtained from 2018:

****

#### Fig 4.9 NMF topic modelling-New topic obtained from 2019:

****

#### Fig 4.10 NMF topic modelling-New topic obtained from 2020:

****