Beyond the Headlines: Machine Learning Insights for News Articles

Team 06 Members: Kevin Murphy, Khushi Jasrapuria, Megha Arul Senthilkumar, Riris Grace Karolina, Shravani Thalla

Dataset & Problem Statement

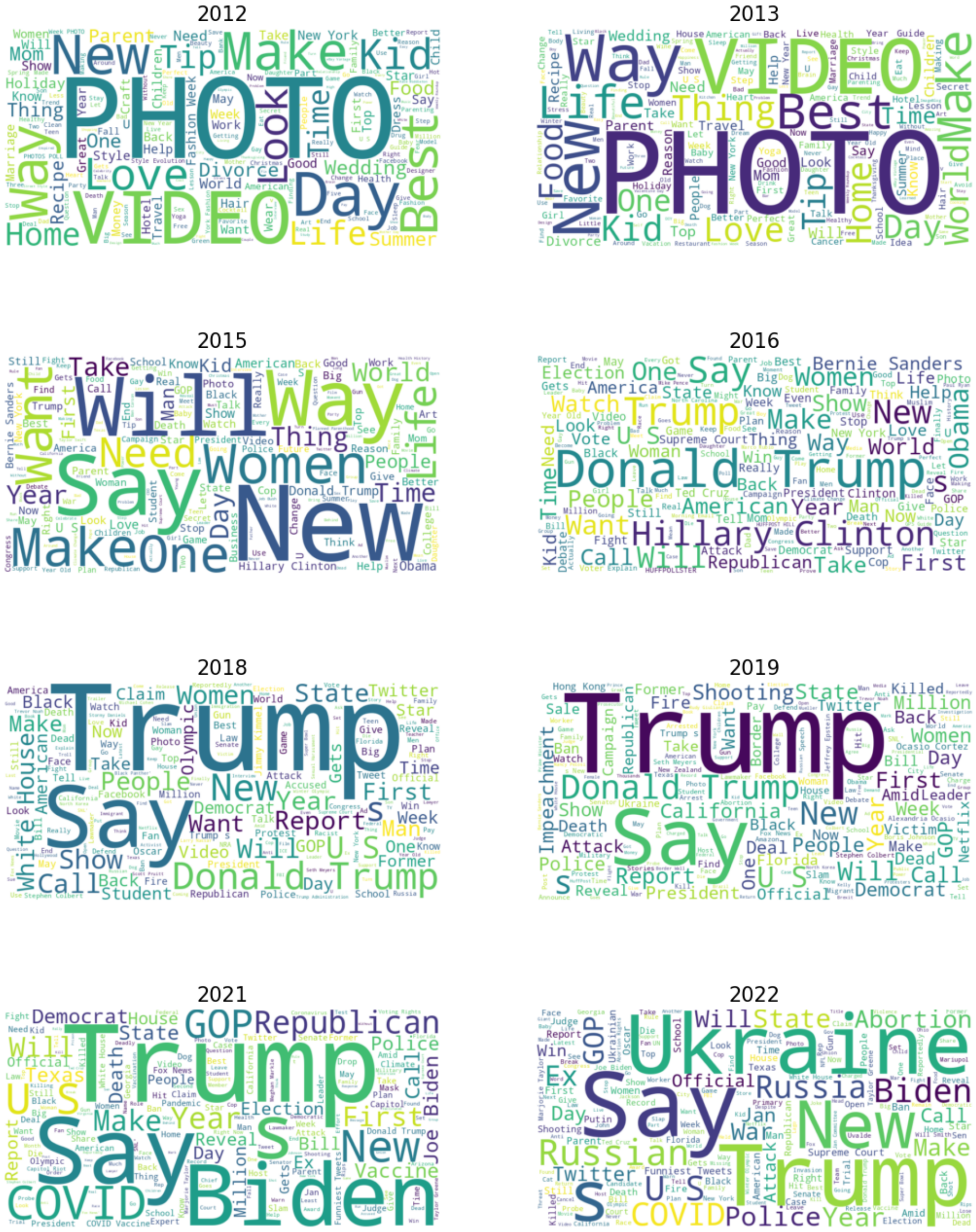
Dataset

HuffPost News 2012-2022. **Rows**: 210k. **Column**: category, headline, authors, link, short description, date. **Label**: 42 categories.

Objective

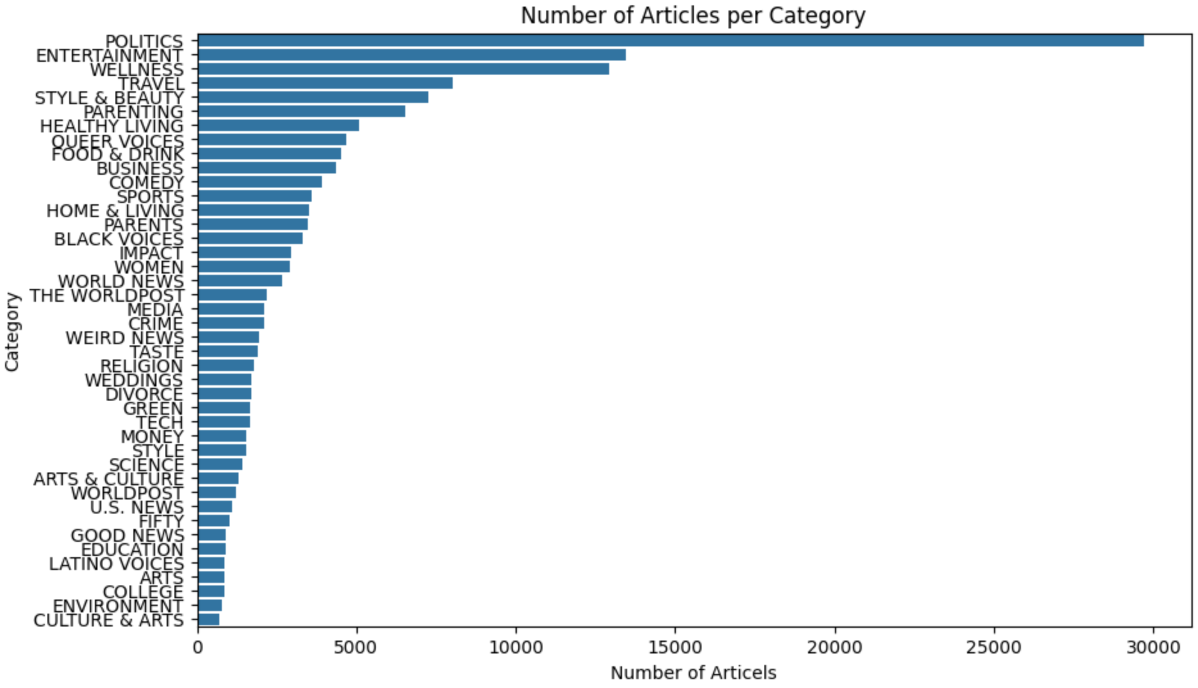
Classify the **news topic** based on similarity of each news category.

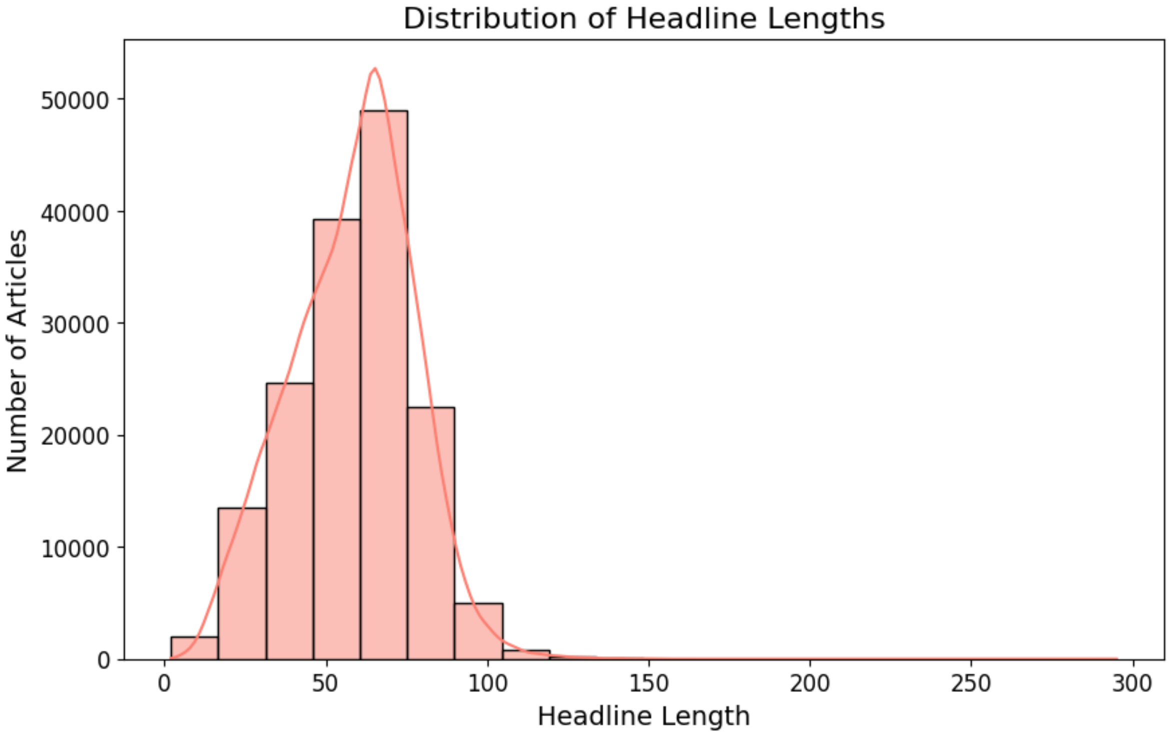
The news categorized into **positive and negative sentiment** to enhance user experience.



Exploratory Data Analysis

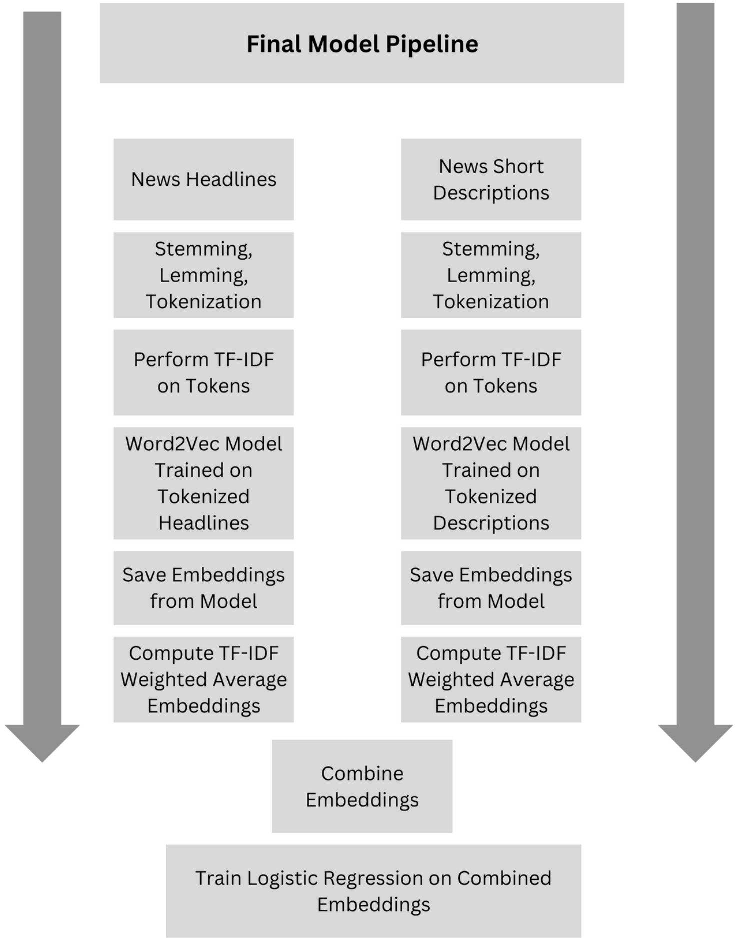
Exploratory Data Analysis





News Category Classification

- Final Model had an accuracy score of 54%
- With 42 unique categories, we feel as though the model produces decent results
- We experimented with BoW, pretrained models, and different classification models
- The best performing model was a logistic regression using Word2Vec models trained on our dataset



Final Model Pipeline

News Headlines

Stemming, Lemming, Tokenization

Perform TF-IDF on Tokens

Word2Vec Model Trained on Tokenized Headlines

Save Embeddings from Model

Compute TF-IDF Weighted Average Embeddings

News Short Descriptions

Stemming, Lemming, Tokenization

Perform TF-IDF on Tokens

Word2Vec Model Trained on Tokenized Descriptions

Save Embeddings from Model

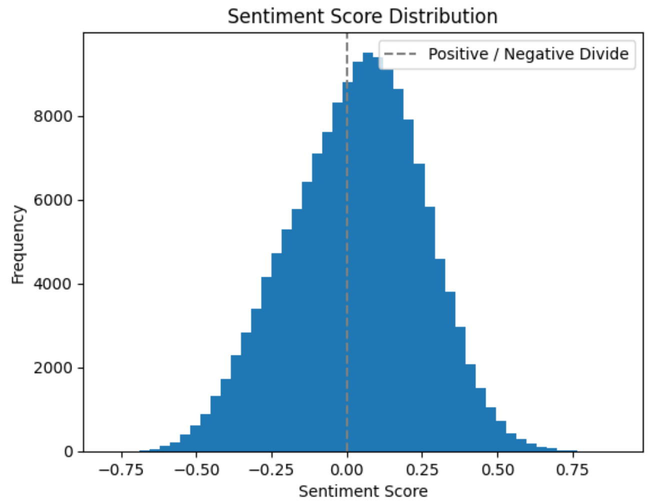
Compute TF-IDF Weighted Average Embeddings

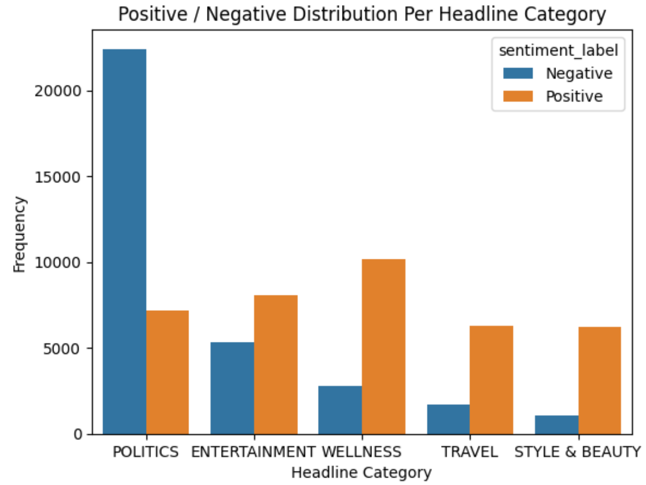
Combine Embeddings

Train Logistic Regression on Combined Embeddings

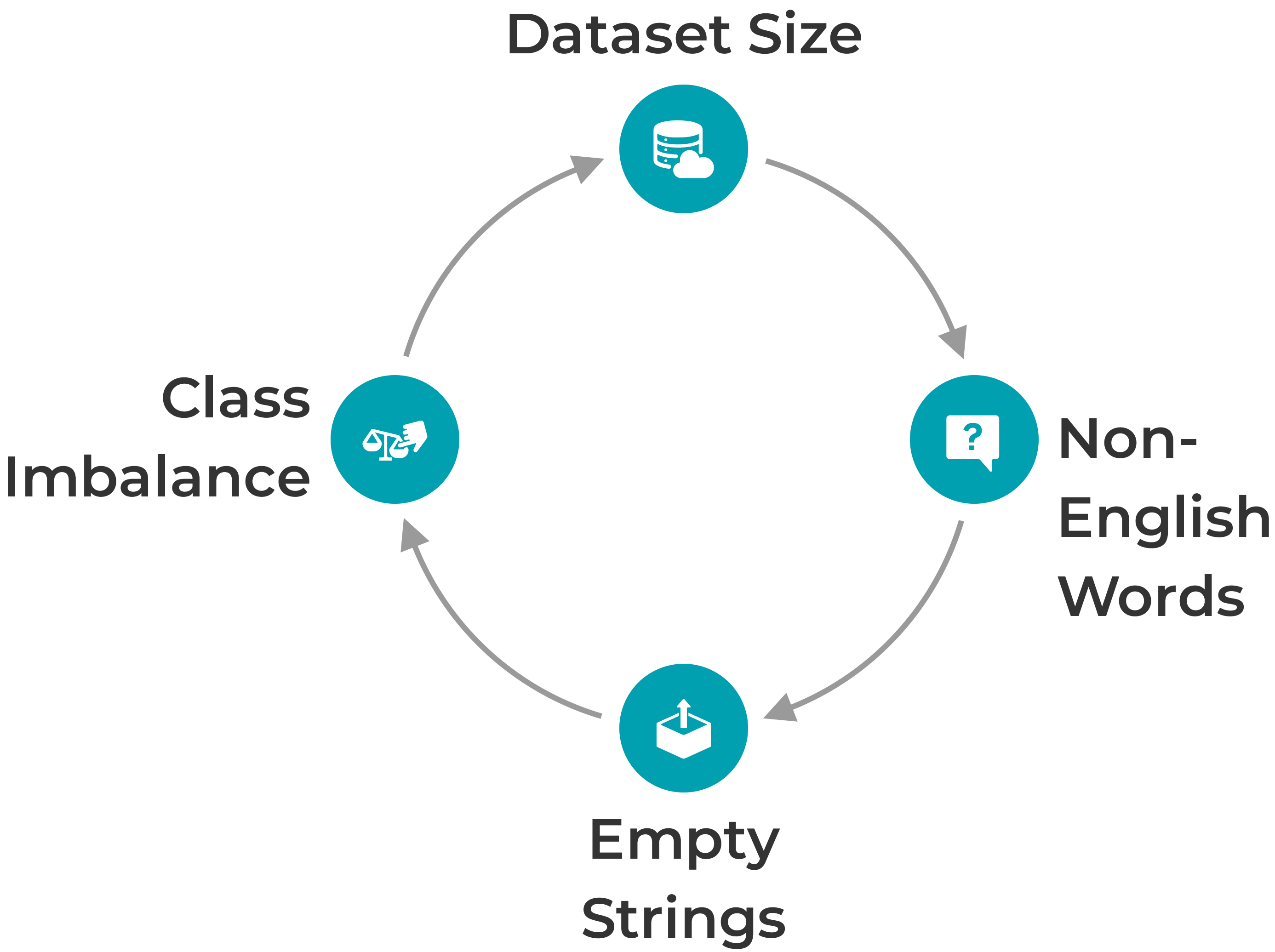
Headline Sentiment Analysis

- Created baseline positive and negative headlines
  - Positive: "breakthrough revolutionary uplifting success innovative pioneering empowering miraculous flourishing renowned"
  - Negative: “catastrophe scandalous devastating oppressive perilous grim corrupt turmoil bankrupt brutal"
- Split at Sentiment Score of 0.0
- Found slightly higher number of positive headlines


Sentiment Score Distribution


Positive / Negative Distribution Per Headline Category

Challenges



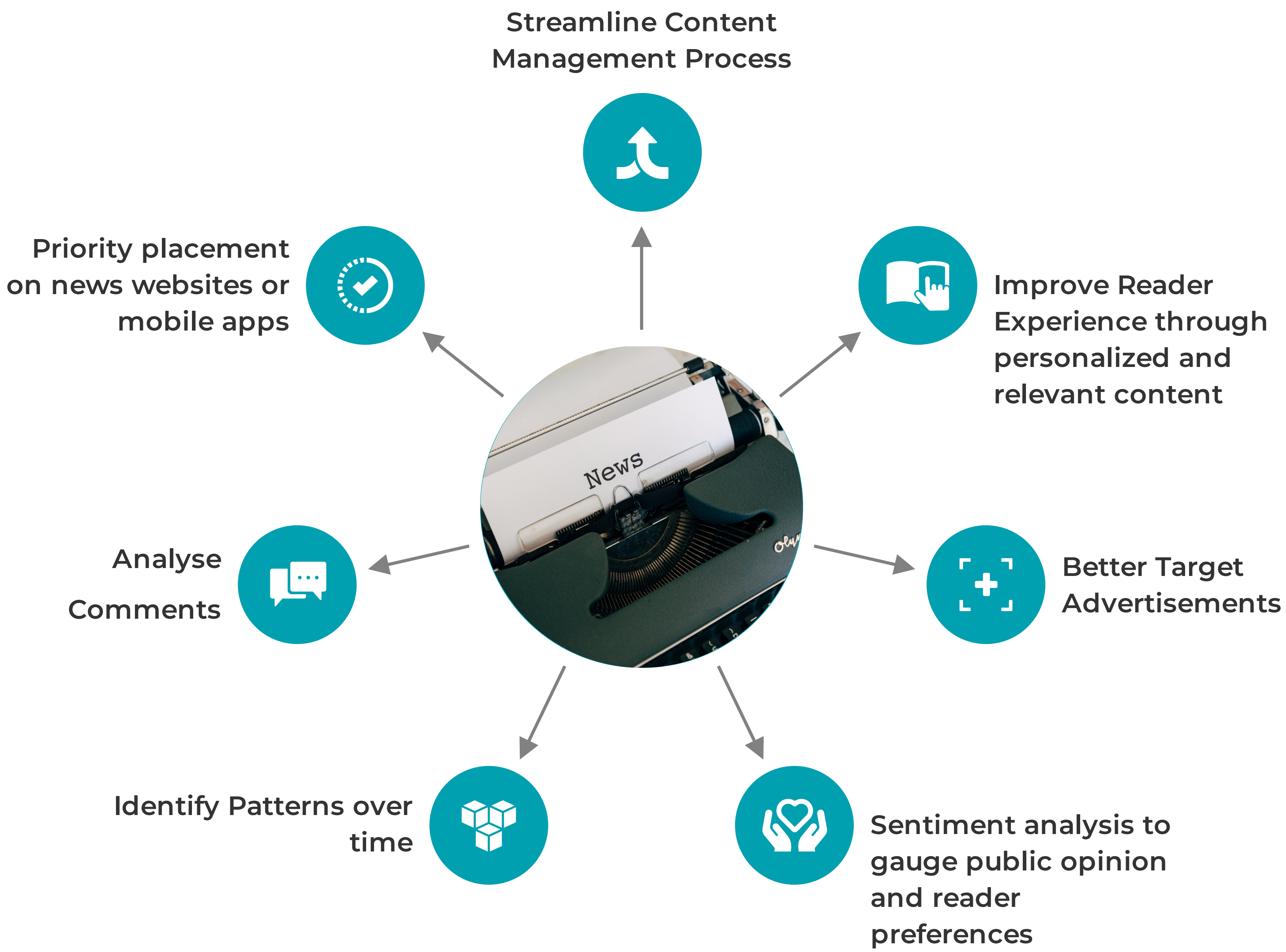
Dataset Size

Non-English Words

Empty Strings

Class Imbalance

Real Word Application



- Streamline Content Management Process
- Improve Reader Experience through personalized and relevant content
- Better Target Advertisements
- Sentiment analysis to gauge public opinion and reader preferences
- Identify Patterns over time
- Analyse Comments
- Priority placement on news websites or mobile apps