NEWS ARTICLE ANALYSIS USING TEXT MINING

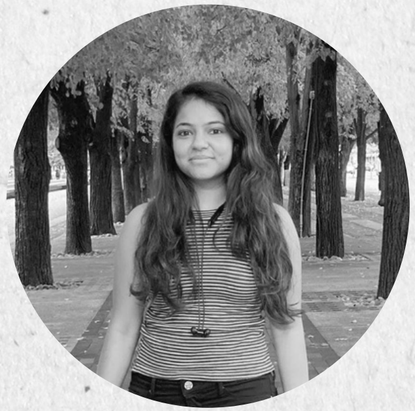
by: Advaith Shyamsunder Rao, Ayush Oturkar, Hsiao-Chun Hung, Vanshita Gupta

MEET THE REPORTERS

Ayush Oturkar

Hsiao-Chun Hung

Advaith Rao



Vanshita Gupta

Introduction and Objectives

- Retrieve news articles from APIs
- Perform data preprocessing to improve data quality
- Use Visualizations to explore data trends
- Topic modeling techniques, such as LDA, cluster the data into topics.
- Further analysis - sentiment analysis and N-gram analysis

Methodology

Web Scraping

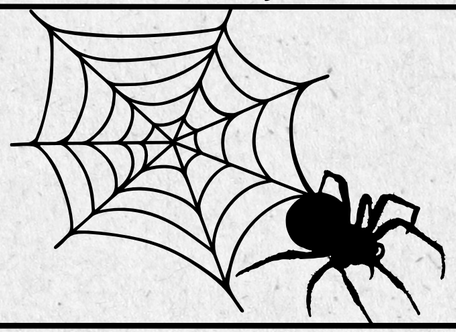
Exploratory Data Analysis

Sentiment Analysis

Data Cleaning

Topic Modeling

Observations, Future Scope



Web Scraping

1. News API
   - Date Range: last 30 days
   - Data Fetch: 1750 rows
   - BBC and CNN News fetched using this API

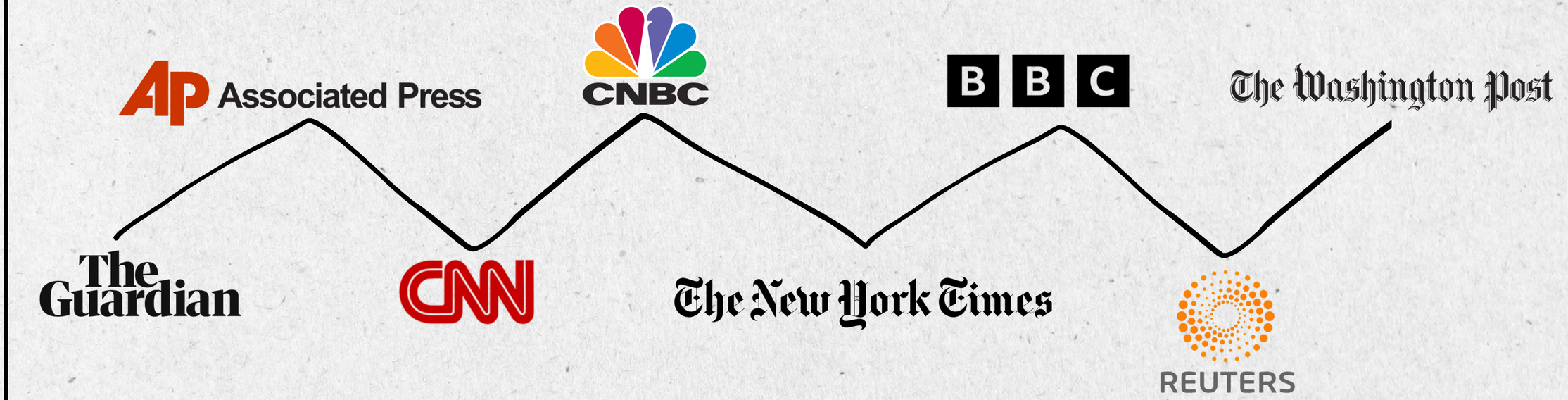
2. CNBC Rapid API
   - Date Range: last 30 days
   - Data Fetch: 50 rows

3. Bing Rapid API
   - Date Range: last 30 days
   - Data Fetch: 757 rows

4. NewsCatcher API
   - Date Range: last 30 days
   - Data Fetch: 3022 rows
   - Reuters, NY Times, The Guardian, The Washington Post, Associated Press news fetched using this API

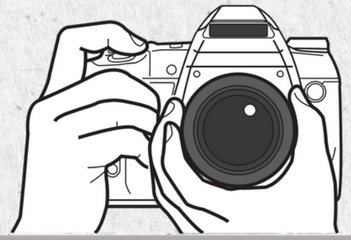
Dataset

News Sources:



Data Timeline: 03-01-2023 to 04-23-2023

Total News Article in Scope: 5,579

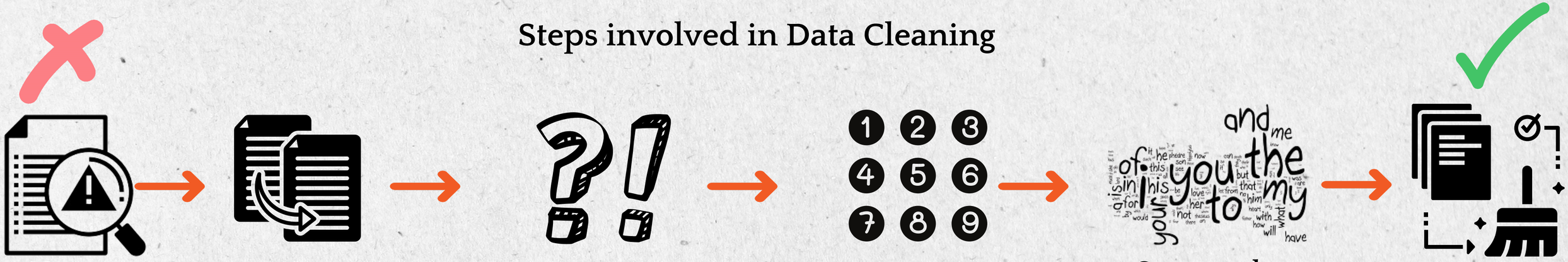


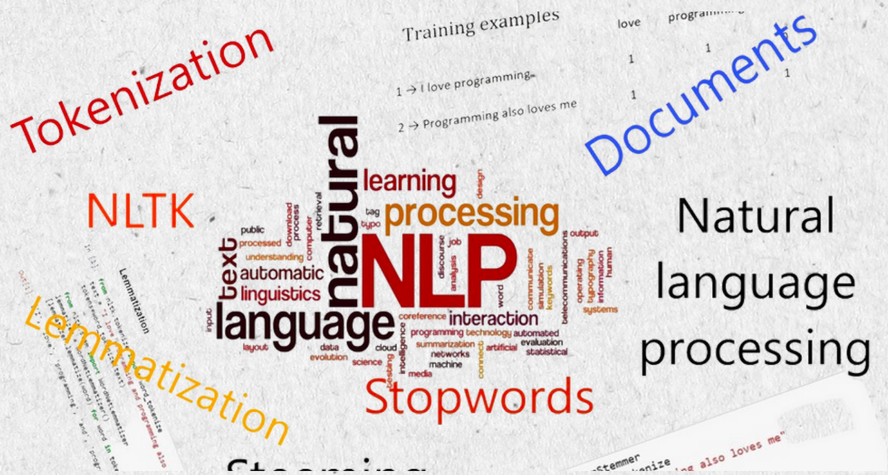
Data Snapshot

article_type	headline	url	description	published_at	news_source
NewsArticle	Why Putin cares about Russia's athletes competing abroad	https://www.bbc.co	While Russia's brutal invasion of	2023-04-16T00:03:50Z	bbc-news
NewsArticle	Al Jaffee: Record-breaking US cartoonist dies at 102	https://www.bbc.co	Award-winning American	2023-04-11T10:01:57Z	bbc-news
NewsArticle	Listen: British Swimming Championships	https://www.bbc.co	British record holder, Ben Proud wil	2023-04-08T16:43:50Z	bbc-news
NewsArticle	Bullying and race bias 'commonplace' in equestrian sports	https://www.bbc.co	The research was commissioned	2023-04-05T23:01:19Z	bbc-news
NewsArticle	EFL coverage set to remain on Sky Sports	https://www.bbc.co	The EFL's current TV deal was a	2023-04-03T15:43:10Z	bbc-news
NewsArticle	100m Olympic champion Fraser-Pryce runs in son's sports day	https://www.bbc.co	Olympic champion Fraser-Pryce run	2023-04-01T11:13:59Z	bbc-news
NewsArticle	WWE & UFC to merge in new sports entertainment brand	https://www.bbc.co	Ronda Rousey wrestles with WWE	2023-04-03T13:19:13Z	bbc-news
NewsArticle	Listen: British Swimming Championships	https://www.bbc.co	British	2023-04-07T13:57:56Z	bbc-news
NewsArticle	Man Utd reach first Women's FA Cup final	https://www.bbc.co	Mary Earps deflected the ball into	2023-04-15T18:10:48Z	bbc-news
NewsArticle	Listen: IPL - Delhi Capitals v Gujarat Titans	https://www.bbc.co	You need one to watch live TV on	2023-04-03T18:21:14Z	bbc-news
NewsArticle	Listen: IPL - Sunrisers Hyderabad v Mumbai Indians	https://www.bbc.co	You need one to watch live TV on	2023-04-17T16:05:47Z	bbc-news
NewsArticle	Australian Grand Prix final practice - radio & text	https://www.bbc.co	It was nice and sunny during the	2023-03-31T02:52:38Z	bbc-news
NewsArticle	Listen: IPL - Rajasthan Royals v Punjab Kings	https://www.bbc.co	You need one to watch live TV on	2023-04-04T18:19:08Z	bbc-news
NewsArticle	Listen: IPL - Delhi Capitals v Mumbai Indians	https://www.bbc.co	You need one to watch live TV on	2023-04-10T15:50:46Z	bbc-news
NewsArticle	Listen: IPL - Punjab Kings v Gujarat Titans	https://www.bbc.co	You need one to watch live TV on	2023-04-12T16:15:13Z	bbc-news
NewsArticle	Australian Grand Prix first practice - radio & text	https://www.bbc.co	Saudi Arabia ended with a bit of a	2023-03-30T08:37:40Z	bbc-news
NewsArticle	Listen: IPL - Royal Challengers Bangalore v Chennai Super Kings	https://www.bbc.co	You need one to watch live TV on	2023-04-16T17:44:27Z	bbc-news
NewsArticle	Listen: IPL - Gujarat Titans v Chennai Super Kings	https://www.bbc.co	You need one to watch live TV on	2023-03-30T10:14:50Z	bbc-news
NewsArticle	'Beale's brutal task - find way to topple merciless Celtic'	https://www.bbc.co	Beale has dropped only two points	2023-04-07T08:10:57Z	bbc-news
NewsArticle	Listen: IPL - Chennai v Lucknow - Stokes, Wood & Moeen play	https://www.bbc.co	You need one to watch live TV on	2023-04-02T15:52:50Z	bbc-news
NewsArticle	Kit and body image 'holding girls back from sport'	https://www.bbc.co	Howard helped England win a first	2023-04-06T23:02:45Z	bbc-news
NewsArticle	Women's FA Cup semi-final: Build-up to Man Utd v Brighton	https://www.bbc.co	Just the one change for	2023-04-14T14:37:39Z	bbc-news
NewsArticle	Listen: IPL - Kolkata Knight Riders v Sunrisers Hyderabad	https://www.bbc.co	You need one to watch live TV on	2023-04-13T17:57:33Z	bbc-news
NewsArticle	Listen: IPL - Royal Challengers Bangalore v Lucknow Super Giants	https://www.bbc.co	You need one to watch live TV on	2023-04-09T16:28:26Z	bbc-news
NewsArticle	Brain injury legal claims group grows to 378 ex-players	https://www.bbc.co	Ex-Wales wing Dafydd James says	2023-04-04T06:44:00Z	bbc-news

Data Cleaning

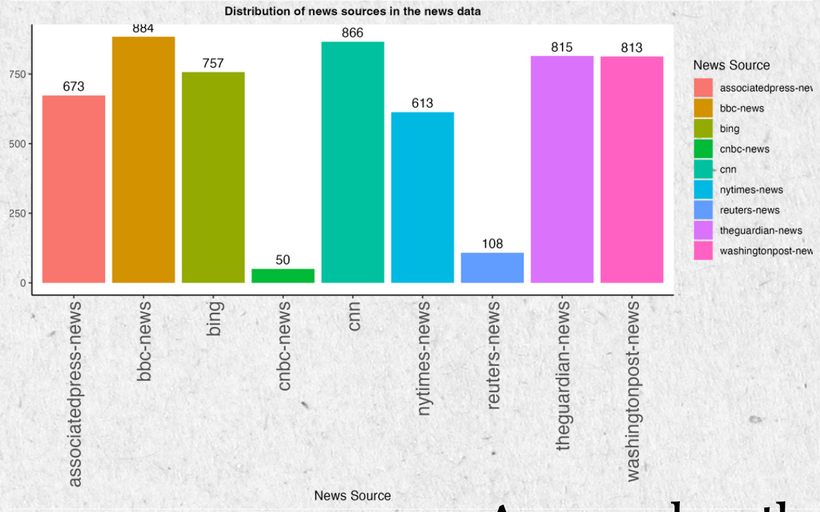
Steps involved in Data Cleaning

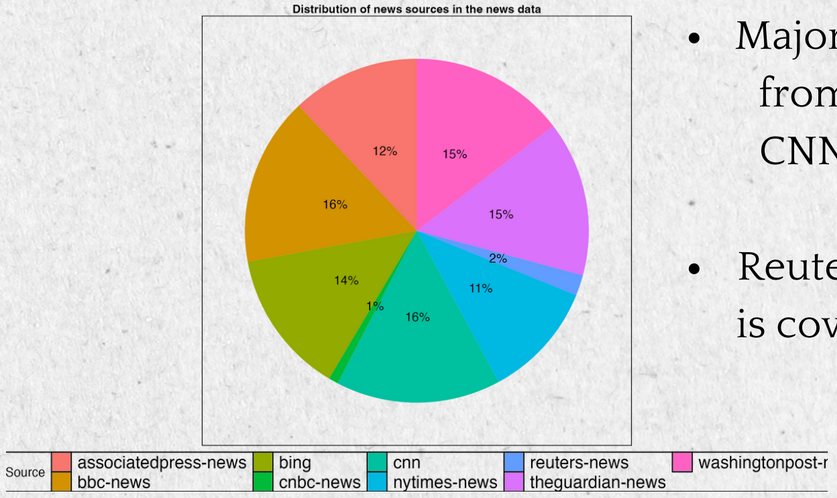




Exploratory Data Analysis

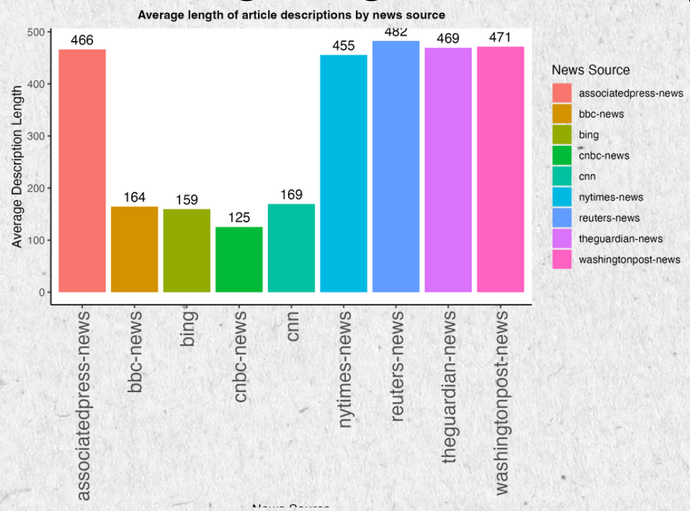
Data split across all news sources





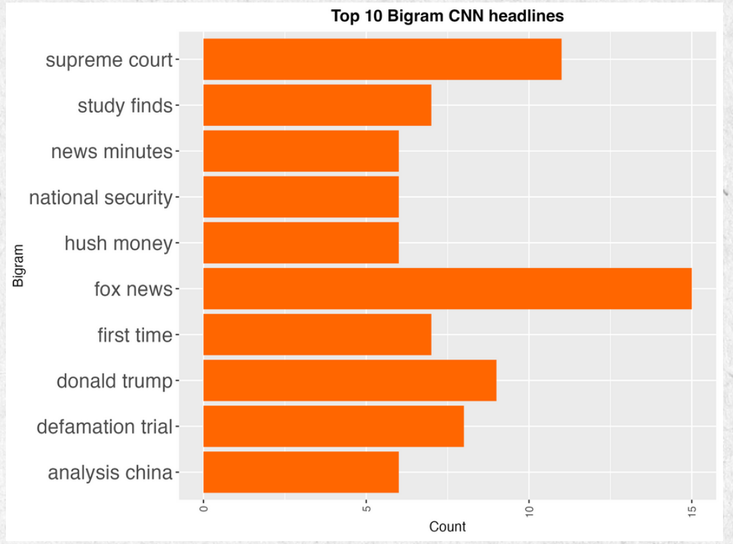
- Majority of news articles from BBC followed by CNN and Washington post
- Reuters and CNBC news is covering just 1% of the data

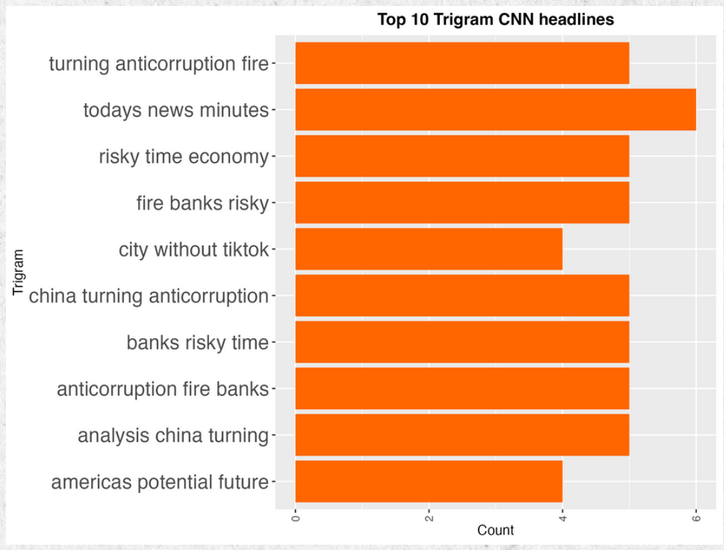
Average length of articles by news sources



- Reuters, NYTimes, Washington post, and the Guardian articles are generally longer in length.
- Bing news articles are generally shorter.

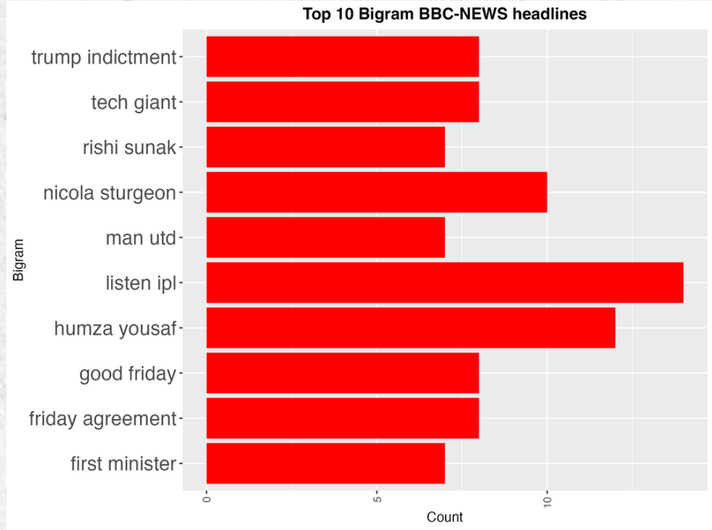
CNN N-Grams Analysis

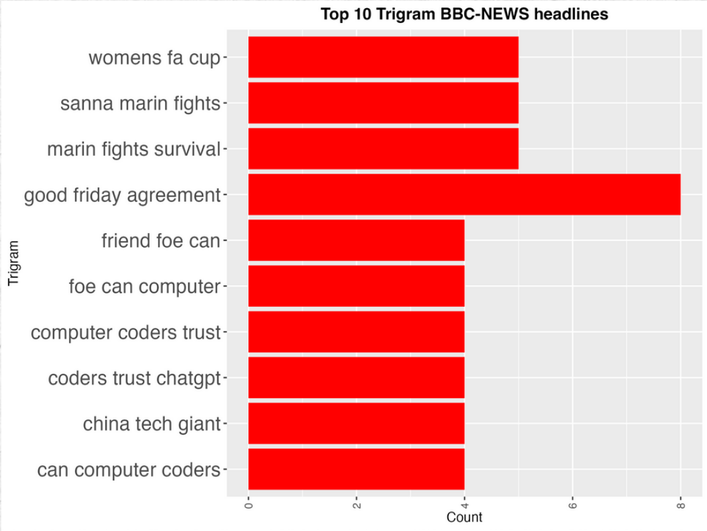




Unigram Wordcloud

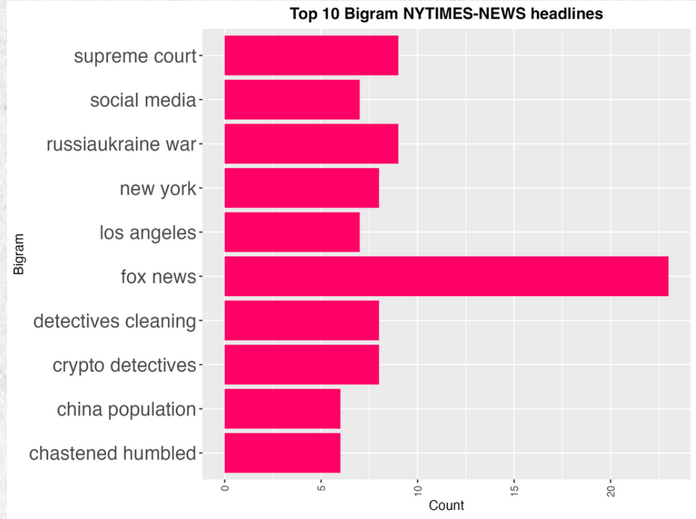
BBC N-Gram Analysis

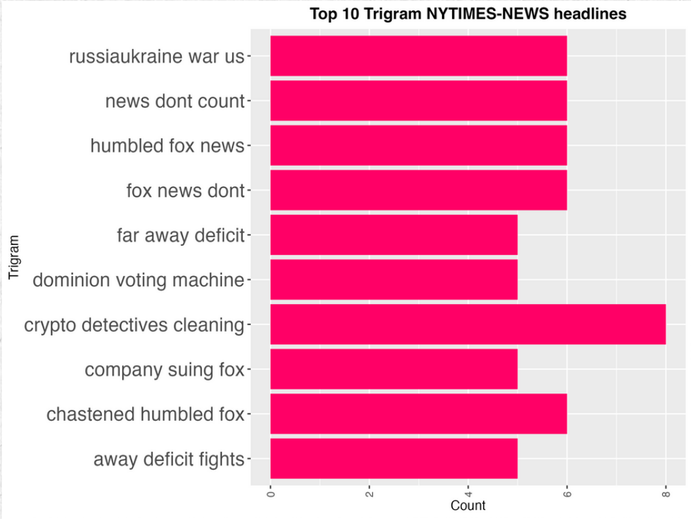




Unigram Wordcloud

New York Times N-gram Analysis





Unigram Wordcloud

Rest of the N-grams from. different sources can be found in the appendix section

Associated Press

CNBC

Washington Post

Reuters

Topic Modeling using LDA

Model used: Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a topic modeling technique used to identify the most likely topics in a corpus of documents, based on the frequency of words that appear in those documents

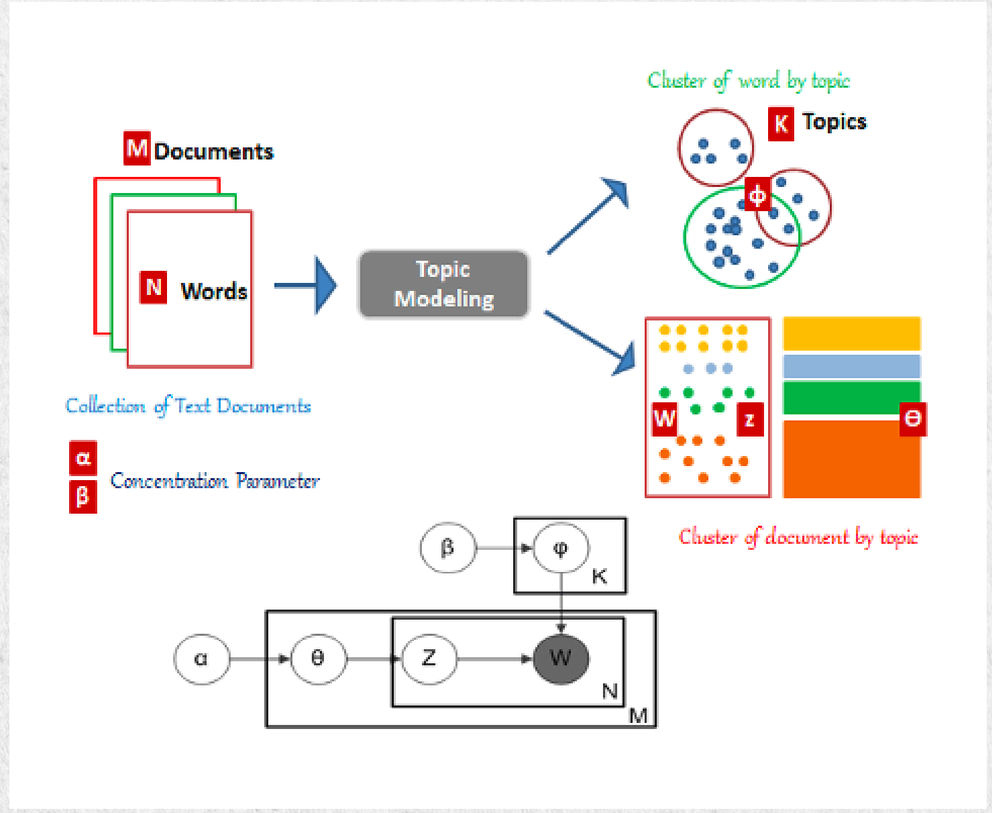
Two Methods of LDA

AEM

AEM is an iterative algorithm that alternates between computing expected sufficient statistics and updating the model's parameters

Gibbs

Gibbs sampling is a Markov Chain Monte Carlo technique that involves iteratively sampling from conditional probability distributions until a stationary distribution is reached

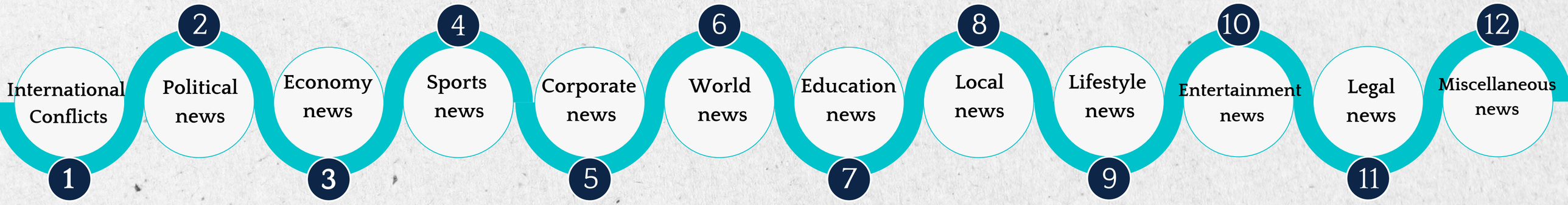


Model Parameters Used:

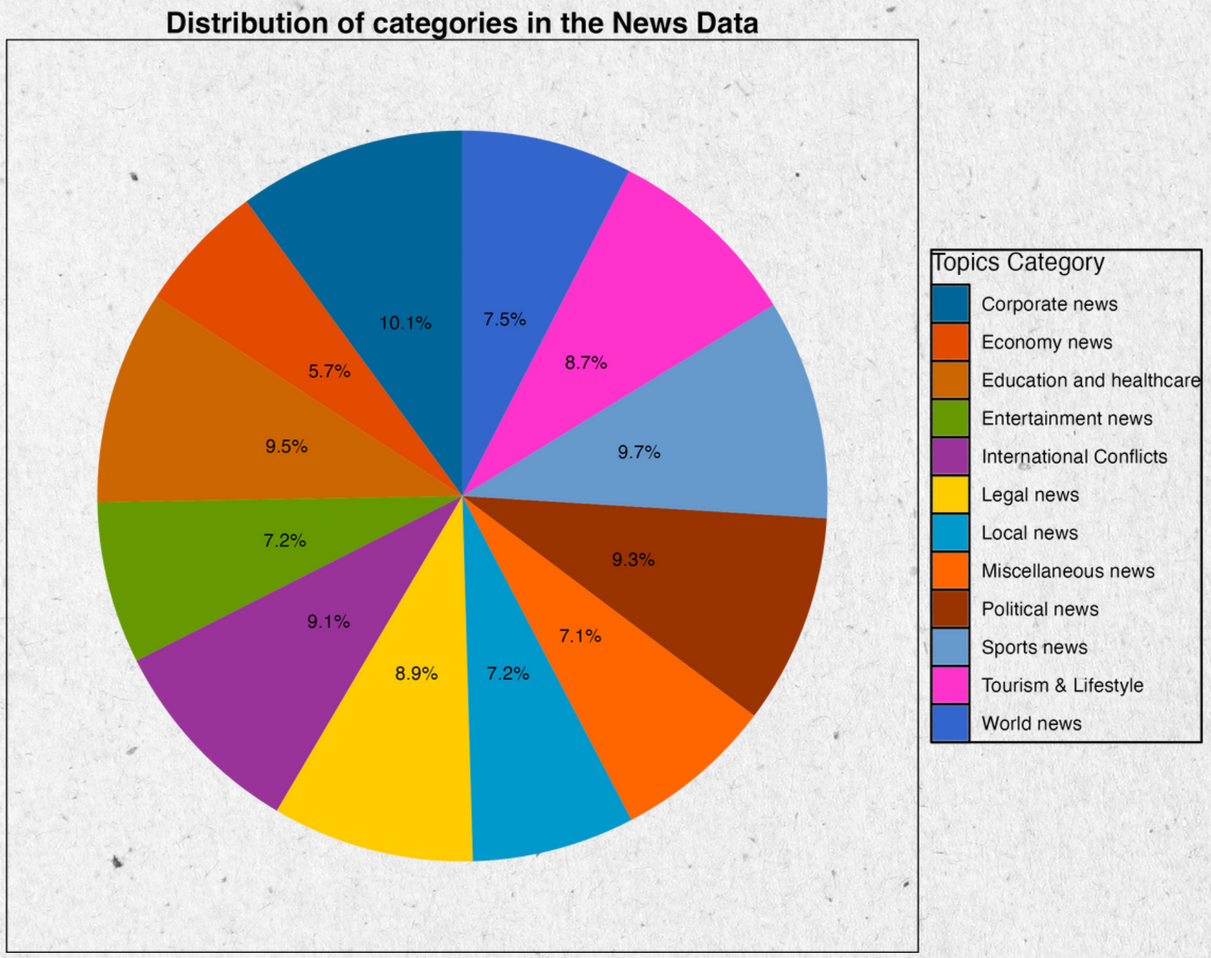
- K (number of topics) : 12 (How?) --Visual Inspection
- Method- Gibbs
- Weights - TFIDF

LDA Output

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
[1,]	"secur"	"char"	"will"	"year"	"compani"	"world"	"state"	"citi"	"like"	"new"	"char"	"comment"
[2,]	"nation"	"minist"	"said"	"sport"	"year"	"one"	"school"	"said"	"can"	"york"	"former"	"share"
[3,]	"ukrain"	"day"	"announc"	"game"	"report"	"new"	"hous"	"home"	"mani"	"work"	"presid"	"stori"
[4,]	"offici"	"first"	"china"	"next"	"busi"	"year"	"bill"	"peopl"	"make"	"show"	"court"	"articl"
[5,]	"russia"	"countri"	"nation"	"team"	"million"	"set"	"educ"	"two"	"say"	"time"	"polit"	"gift"
[6,]	"war"	"govern"	"unit"	"major"	"bank"	"use"	"republican"	"polic"	"way"	"give"	"trump"	"news"
[7,]	"said"	"parti"	"presid"	"leagu"	"accord"	"open"	"right"	"offic"	"one"	"televis"	"feder"	"get"
[8,]	"group"	"leader"	"govern"	"will"	"month"	"across"	"univers"	"charg"	"peopl"	"april"	"washington"	"plan"
[9,]	"russian"	"prime"	"monday"	"season"	"increas"	"australia"	"say"	"kill"	"chang"	"book"	"biden"	"fox"
[10,]	"social"	"forc"	"group"	"play"	"last"	"india"	"law"	"yearold"	"time"	"live"	"donald"	"listen"
[11,]	"militari"	"sinc"	"union"	"three"	"tax"	"ago"	"support"	"arrest"	"around"	"star"	"campaign"	"elect"
[12,]	"includ"	"fight"	"thursday"	"final"	"financi"	"last"	"health"	"man"	"see"	"look"	"hous"	"vote"
[13,]	"latest"	"sudan"	"trade"	"time"	"number"	"place"	"student"	"author"	"now"	"imag"	"investig"	"min"
[14,]	"defens"	"continu"	"econom"	"last"	"cut"	"just"	"public"	"death"	"good"	"last"	"editor"	"sign"
[15,]	"media"	"capit"	"repres"	"first"	"price"	"becom"	"governor"	"includ"	"want"	"age"	"alleg"	"experi"



Category Distribution in Data



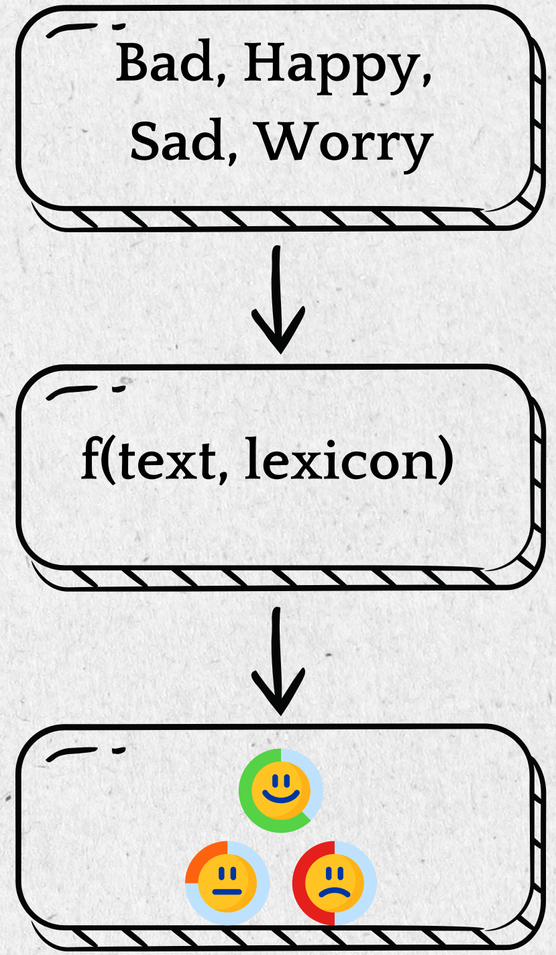
- The news articles are distributed fairly evenly across all categories, with each category accounting for roughly the same percentage of the total articles.
- Roughly 10% of the news articles are clustered as corporate news.
- Close to 9.7% of the news articles are segregated as sports news.
- Least percentage share of 5.7% is taken up by Miscellaneous news which comprises of editorials, opinion pieces, and human interest stories.

Sentiment Analysis

Performed Sentiment Analysis on news headlines using Bing and Afinn Lexicon

Bing

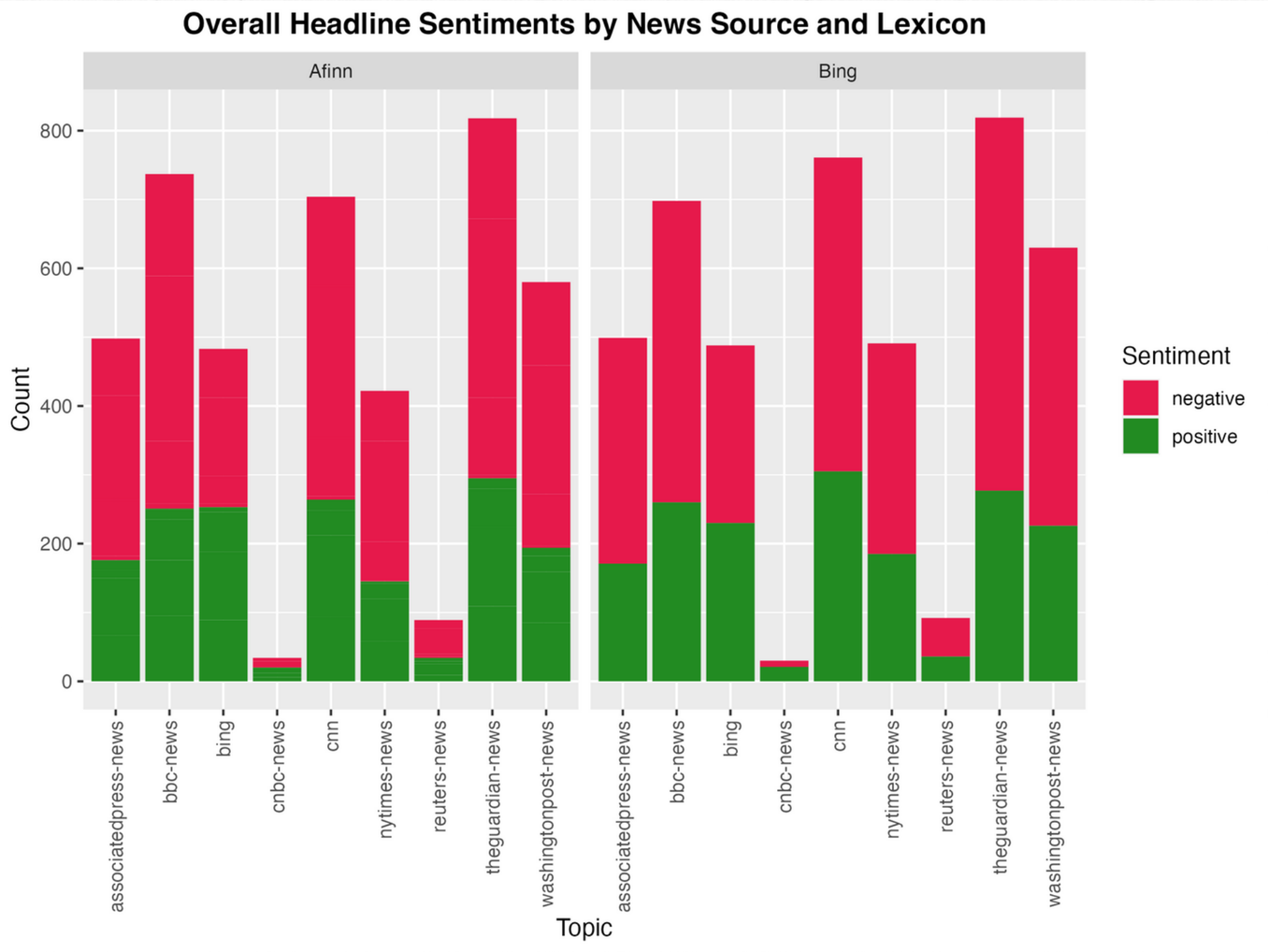
- The Bing Lexicon is a pre-built sentiment analysis tool provided by Microsoft's Bing search engine.
- The score ranges from -1 (most negative) to +1 (most positive), with 0 indicating a neutral sentiment.
- It uses a large lexicon of words and phrases with assigned sentiment scores to analyze the sentiment of text.
- The sentiment score for a given text is calculated based on the number and magnitude of positive and negative words in the lexicon that appear in the text.



Afinn

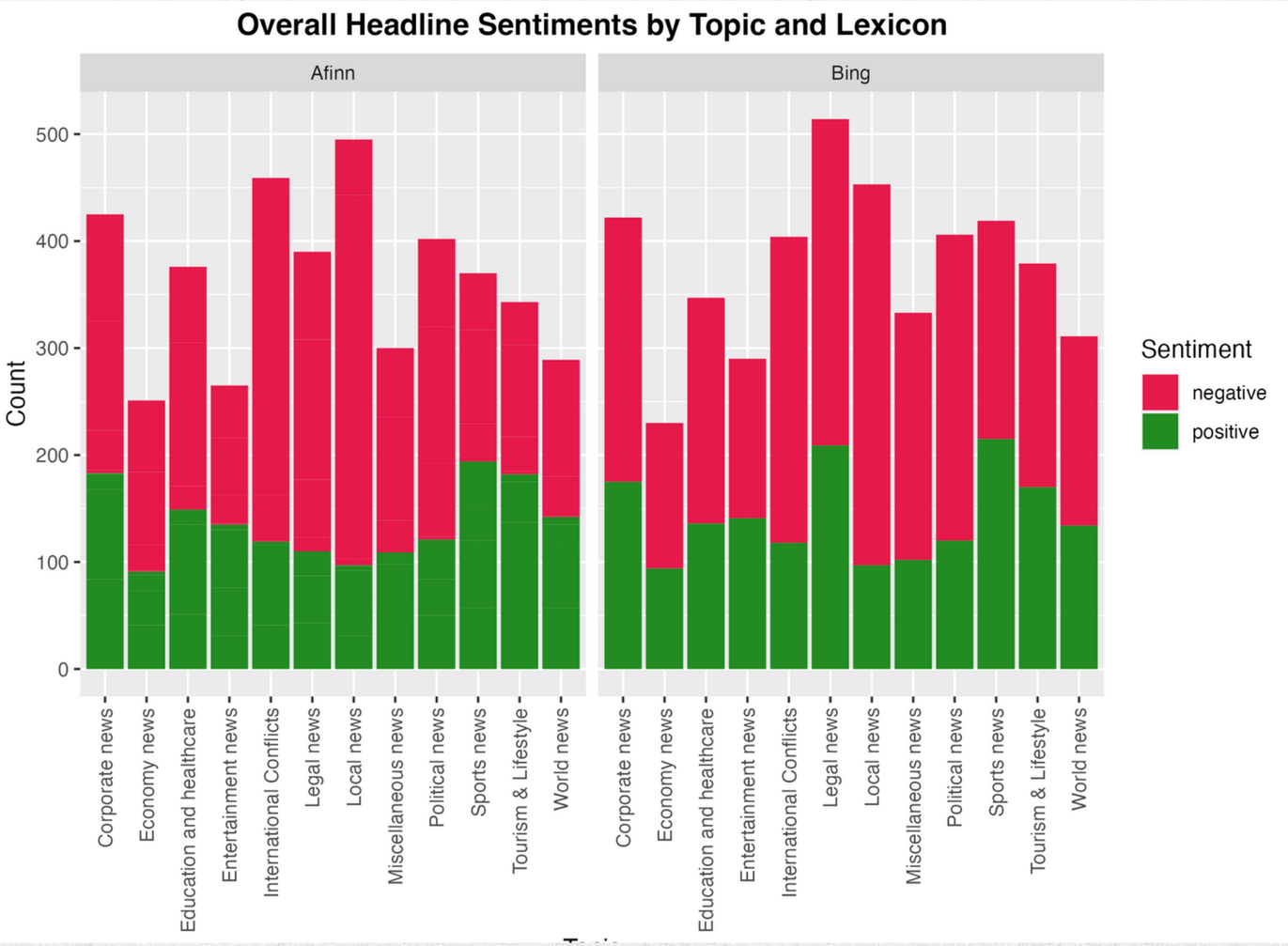
- Afinn Lexicon is a list of words rated for valence or sentiment, with a score ranging from -5 to 5.
- Sentiment analysis using Afinn Lexicon involves assigning a sentiment score to each word in a text, and then aggregating these scores to calculate an overall sentiment score for the text.
- One limitation of Afinn Lexicon is that it may not capture the nuances of language and context, leading to inaccuracies in sentiment analysis.

Sentiment Analysis Output - By News Source



- During the last 30 days of data the news sentiments have been negative across majority of news sources.
- Exceptionally the proportion of positive sentiments in CNBC appears to be notably higher.
- News from Bing looks more balanced

Sentiment Analysis Output - By Topic



- Most of the news we can observe in each category carry negative sentiments except sports and tourism according to both Afinn and Bing Lexicon
- Local news have majority of negative sentiments news articles according to both Afinn and Bing
- Afinn marks majority of legal news as negative while the this is not the case with Bing lexicon

Future Scope of Work

- Wrangle for more data
- Modeling techniques such as NMF, BertTopic
- Named Entity Recognition
- Breaking News Detection
- Fake News Detection

Conclusion

It is evident that through text mining of news articles, we can harness a lot of information on various aspects of news reporting, including the tone, sentiment, and topics covered.

By examining a large sample of news articles, it is possible to identify trends and patterns that may not be apparent through manual analysis alone.

BREAKING NEWS

The END might be here!

- Team 8