In [50]:

```
import matplotlib as plt
import matplotlib.pyplot as plt
```

In [51]:

```
import pandas as pd
import os , sys
import yaml
notebook_dir = os.getcwd()

sys.path.append(os.path.abspath(os.path.join(notebook_dir, '..')))
```

In [1]:

```
import sys
import os
import pandas as pd

# Add the 'scripts' directory to the Python path
sys.path.append(os.path.abspath('../scripts'))
```

In [58]:

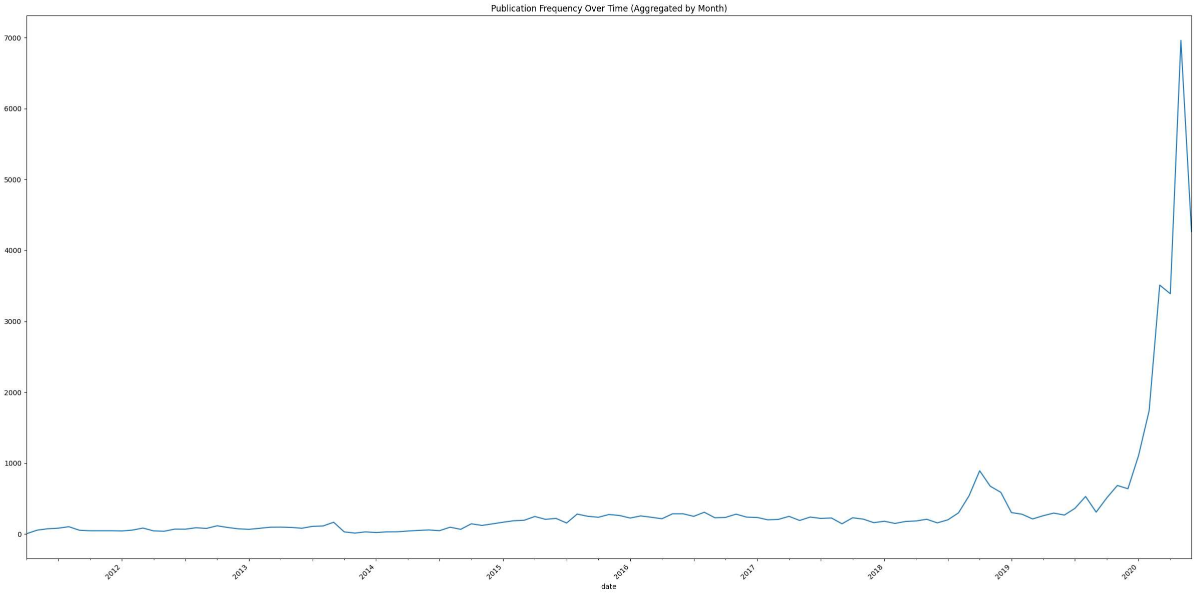
```
from data_loader import load_data, inspect_data, handle_missing_values, load_config

# Load configuration with the correct path
config = load_config(config_file='../config.yaml')

# Step 1: Load Dataset
data = load_data(config['dataset_path'])
print (data.columns)
```

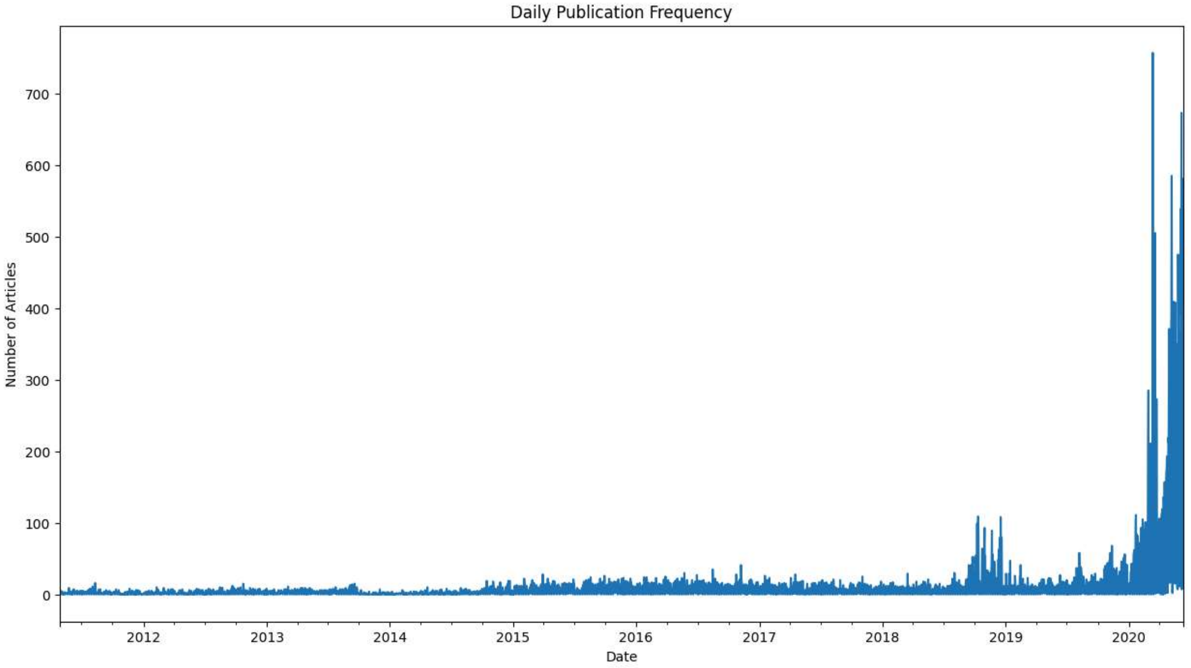
```
Dataset loaded successfully with 1048575 rows and 6 columns.
Index(['Unnamed: 0', 'headline', 'url', 'publisher', 'date', 'stock'], dtype='obje
ct')
```

In [54]:

```
#Time analysis
from time_series_analysis import publication_frequency

# Publication Frequency Analysis
#publication_frequency(data_cleaned)

#dataT = pd.DataFrame({'date': pd.date_range(start='2020-01-01', periods=3650, freq
publication_frequency(data)

# Daily Publication Frequency

#data['date'] = pd.to_datetime(data['date'], errors='coerce')

# Set the 'date' column as the index
data.set_index('date', inplace=True)

# Resample the data by day to get the count of articles per day
daily_publications = data['headline'].resample('D').count()

# Plot the publication frequency over time
plt.figure(figsize=(15, 8))
daily_publications.plot()
plt.title('Daily Publication Frequency')
plt.xlabel('Date')
plt.ylabel('Number of Articles')
plt.show()
```

```
c:\Users\user\Desktop\KIFIYA Projects\Nova-Financial-Solutions-Week-01\scripts\tim
e_series_analysis.py:9: UserWarning: The argument 'infer_datetime_format' is depre
cated and will be removed in a future version. A strict version of it is now the d
efault, see https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.ht
ml. You can safely remove this argument.
  data['date'] = pd.to_datetime(data['date'], infer_datetime_format=True, errors
='coerce')
c:\Users\user\Desktop\KIFIYA Projects\Nova-Financial-Solutions-Week-01\scripts\tim
e_series_analysis.py:15: UserWarning: Converting to PeriodArray/Index representati
on will drop timezone information.
  freq = data['date'].dt.year.value_counts().sort_index()
```





In [55]:

```
import importlib
import visualizers
importlib.reload(visualizers)
```

Out[55]:

```
<module 'visualizers' from 'c:\\Users\\user\\Desktop\\KIFIYA Projects\\Nova-Financ
ial-Solutions-Week-01\\scripts\\visualizers.py'>
```

In [57]:

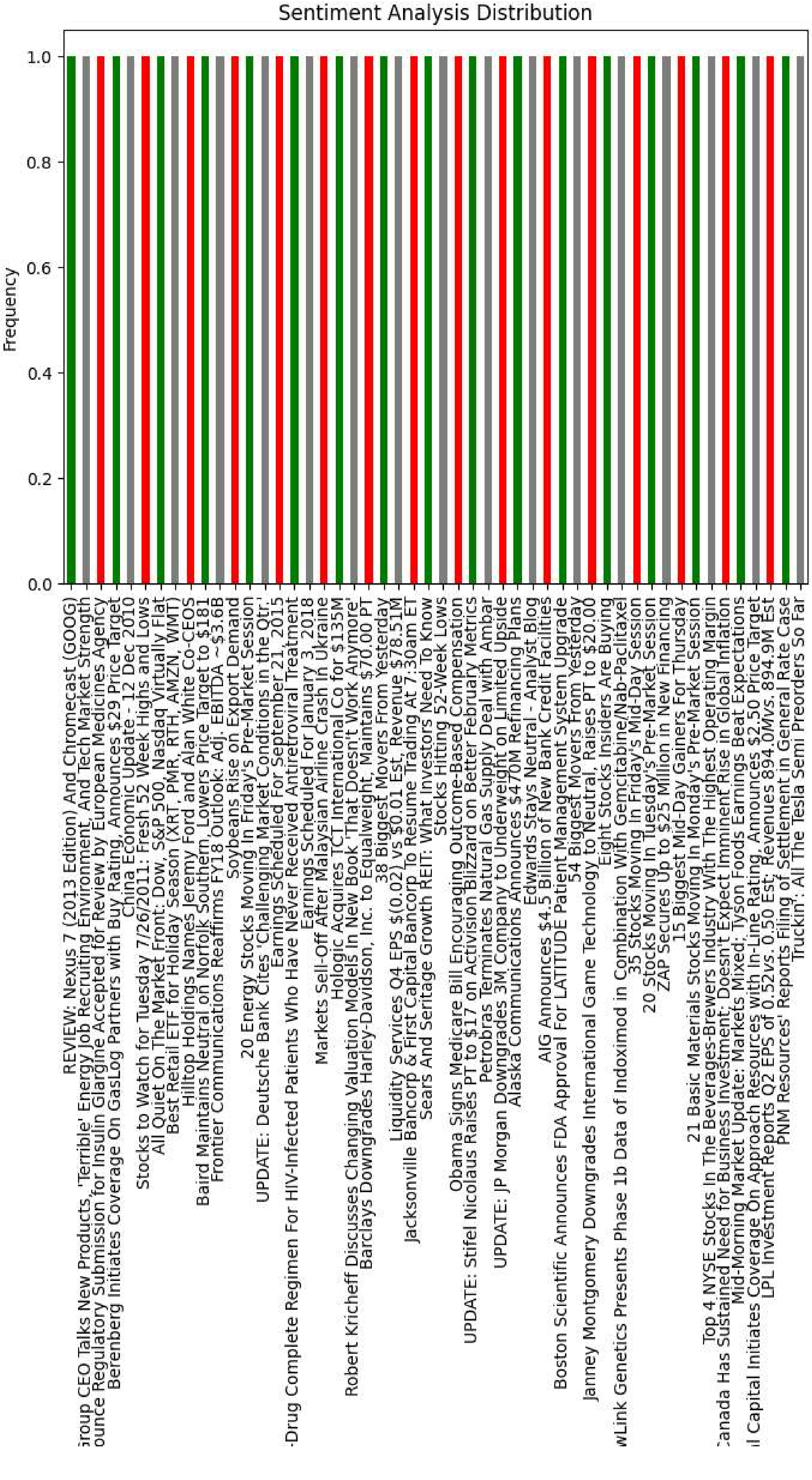
```
print (data.columns)
```

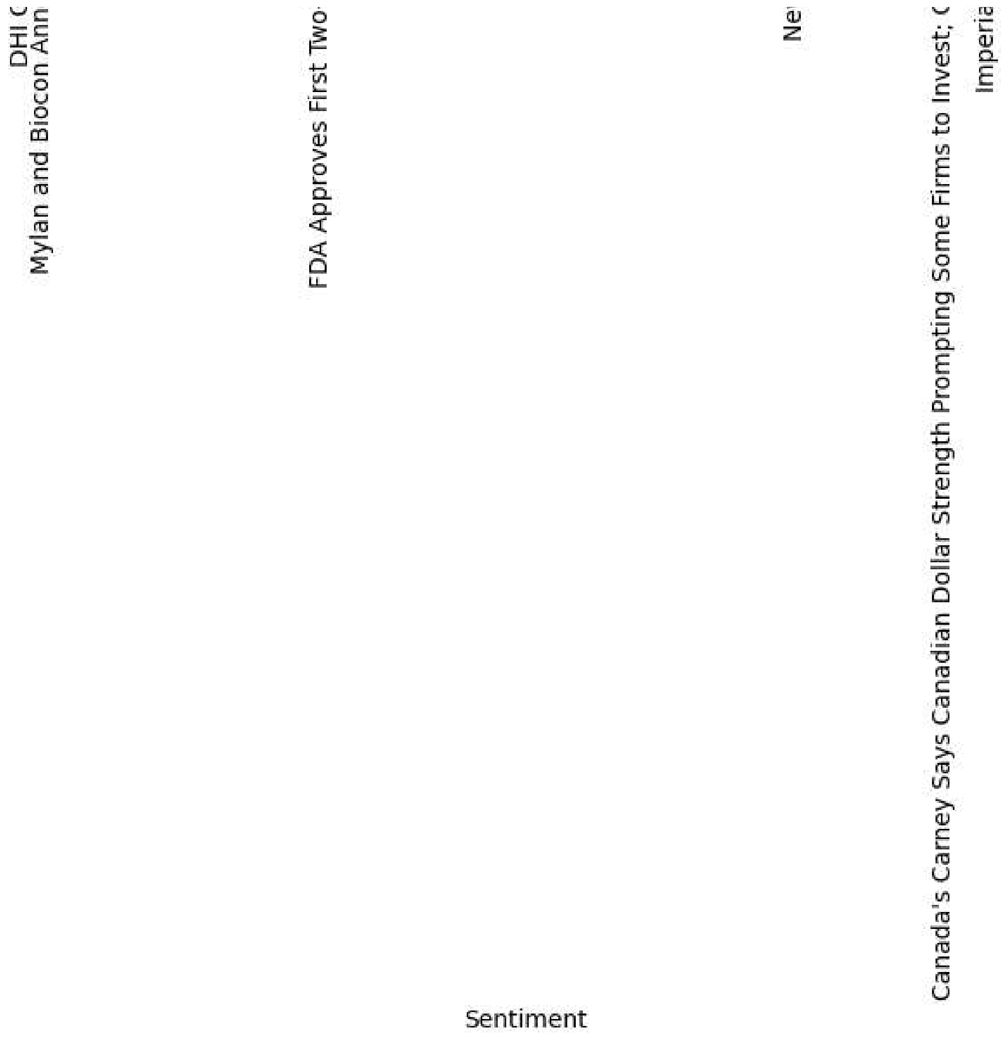
```
Index(['Unnamed: 0', 'headline', 'url', 'publisher', 'stock'], dtype='object')
```

In [60]:

```
# Visualize task-1 analysis
from Task1.visualizers import (
```

```python
    plot_sentiment_analysis,
    plot_time_series,
    plot_publisher_contribution,
    plot_word_cloud,
    plot_stock_article_count,
    plot_word_cloud_top_stocks
)

sample_size = 50  # Specify the number of random samples you want

plot_sentiment_analysis(data, sentiment_column='headline', sample_size=sample_size)
plot_word_cloud(data, text_column='headline', sample_size=sample_size)
plot_time_series(data, date_col='date', sample_size=sample_size)
plot_publisher_contribution(data, publisher_column='publisher', top_n=5, sample_siz
plot_stock_article_count(data, stock_column='stock', sample_size=sample_size)
plot_word_cloud_top_stocks(data, text_column='headline', stock_column='stock', top_
```

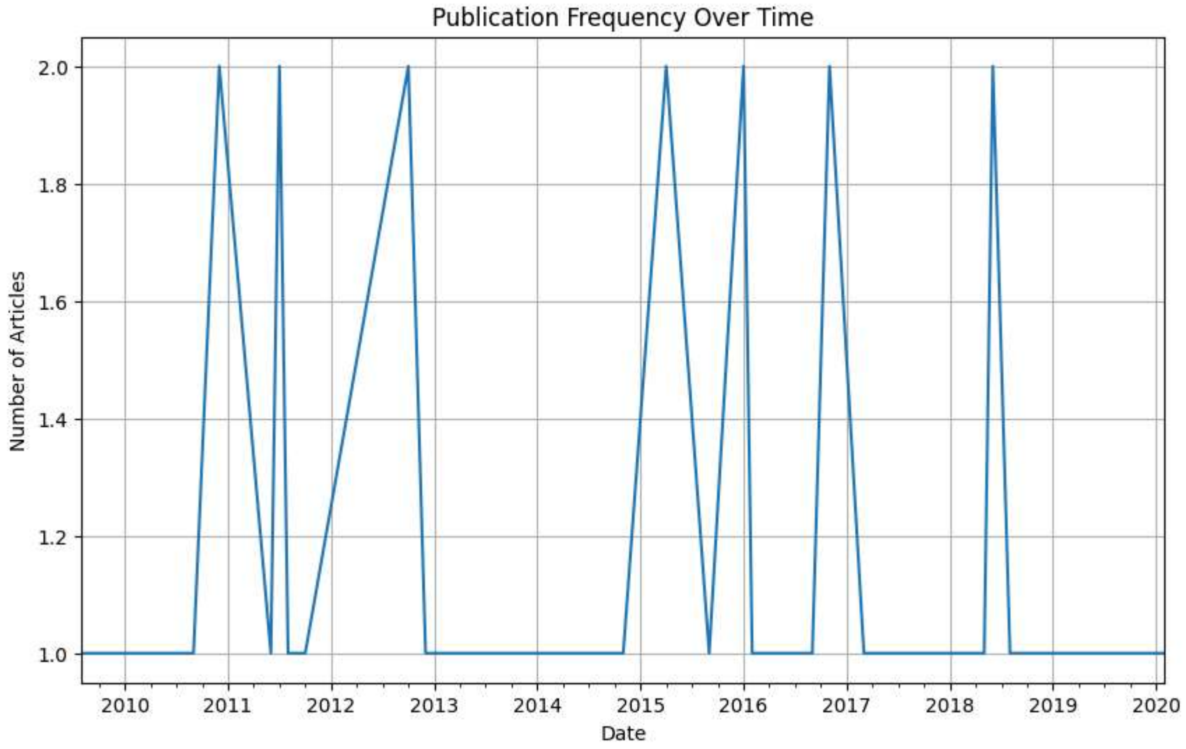
Sentiment Analysis Distribution
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
REVIEW: Nexus 7 (2013 Edition) And Chromecast (GOOG)
roup CEO Talks New Products, 'Terrible' Energy Job Recruiting Environment, And Tech Market Strength
ounce Regulatory Submission for Insulin Glargine Accepted for Review by European Medicines Agency
Berenberg Initiates Coverage On GasLog Partners with Buy Rating, Announces $29 Price Target
China Economic Update - 12 Dec 2010
Stocks to Watch for Tuesday 7/26/2011: Fresh 52 Week Highs and Lows
All Quiet On The Market Front: Dow, S&P 500, Nasdaq Virtually Flat
Best Retail ETF for Holiday Season (XRT, PMR, RTH, AMZN, WMT)
Hilltop Holdings Names Jeremy Ford and Alan White Co-CEOs
Baird Maintains Neutral on Norfolk Southern, Lowers Price Target to $181
Frontier Communications Reaffirms FY18 Outlook: Adj. EBITDA ~$3.6B
Soybeans Rise on Export Demand
20 Energy Stocks Moving In Friday's Pre-Market Session
UPDATE: Deutsche Bank Cites 'Challenging Market Conditions in the Qtr.'
Earnings Scheduled For September 21, 2015
-Drug Complete Regimen For HIV-Infected Patients Who Have Never Received Antiretroviral Treatment
Earnings Scheduled For January 3, 2018
Markets Sell-Off After Malaysian Airline Crash In Ukraine
Hologic Acquires TCT International Co for $135M
Robert Kricheff Discusses Changing Valuation Models In New Book 'That Doesn't Work Anymore'
Barclays Downgrades Harley-Davidson, Inc. to Equalweight, Maintains $70.00 PT
38 Biggest Movers From Yesterday
Liquidity Services Q4 EPS $(0.02) vs $0.01 Est, Revenue $78.51M
Jacksonville Bancorp & First Capital Bancorp To Resume Trading At 7:30am ET
Sears And Seritage Growth REIT: What Investors Need To Know
Stocks Hitting 52-Week Lows
Obama Signs Medicare Bill Encouraging Outcome-Based Compensation
UPDATE: Stifel Nicolaus Raises PT to $17 on Activision Blizzard on Better February Metrics
Petrobras Terminates Natural Gas Supply Deal with Ambar
UPDATE: JP Morgan Downgrades 3M Company to Underweight on Limited Upside
Alaska Communications Announces $470M Refinancing Plans
Edwards Stays Neutral - Analyst Blog
AIG Announces $4.5 Billion of New Bank Credit Facilities
Boston Scientific Announces FDA Approval For LATITUDE Patient Management System Upgrade
54 Biggest Movers From Yesterday
Janney Montgomery Downgrades International Game Technology to Neutral, Raises PT to $20.00
Eight Stocks Insiders Are Buying
wLink Genetics Presents Phase 1b Data of Indoximod in Combination With Gemcitabine/Nab-Paclitaxel
35 Stocks Moving In Friday's Mid-Day Session
20 Stocks Moving In Tuesday's Pre-Market Session
ZAP Secures Up to $25 Million in New Financing
15 Biggest Mid-Day Gainers For Thursday
21 Basic Materials Stocks Moving In Monday's Pre-Market Session
Top 4 NYSE Stocks In The Beverages-Brewers Industry With The Highest Operating Margin
anada Has Sustained Need for Business Investment; Doesn't Expect Imminent Rise in Global Inflation
Mid-Morning Market Update: Markets Mixed; Tyson Foods Earnings Beat Expectations
l Capital Initiates Coverage On Approach Resources with In-Line Rating, Announces $2.50 Price Target
LPL Investment Reports Q2 EPS of 0.52vs. 0.50 Est; Revenues 894.0Mvs. 894.9M Est
PNM Resources' Reports Filing of Settlement in General Rate Case
Truckin': All The Tesla Semi Preorders So Far

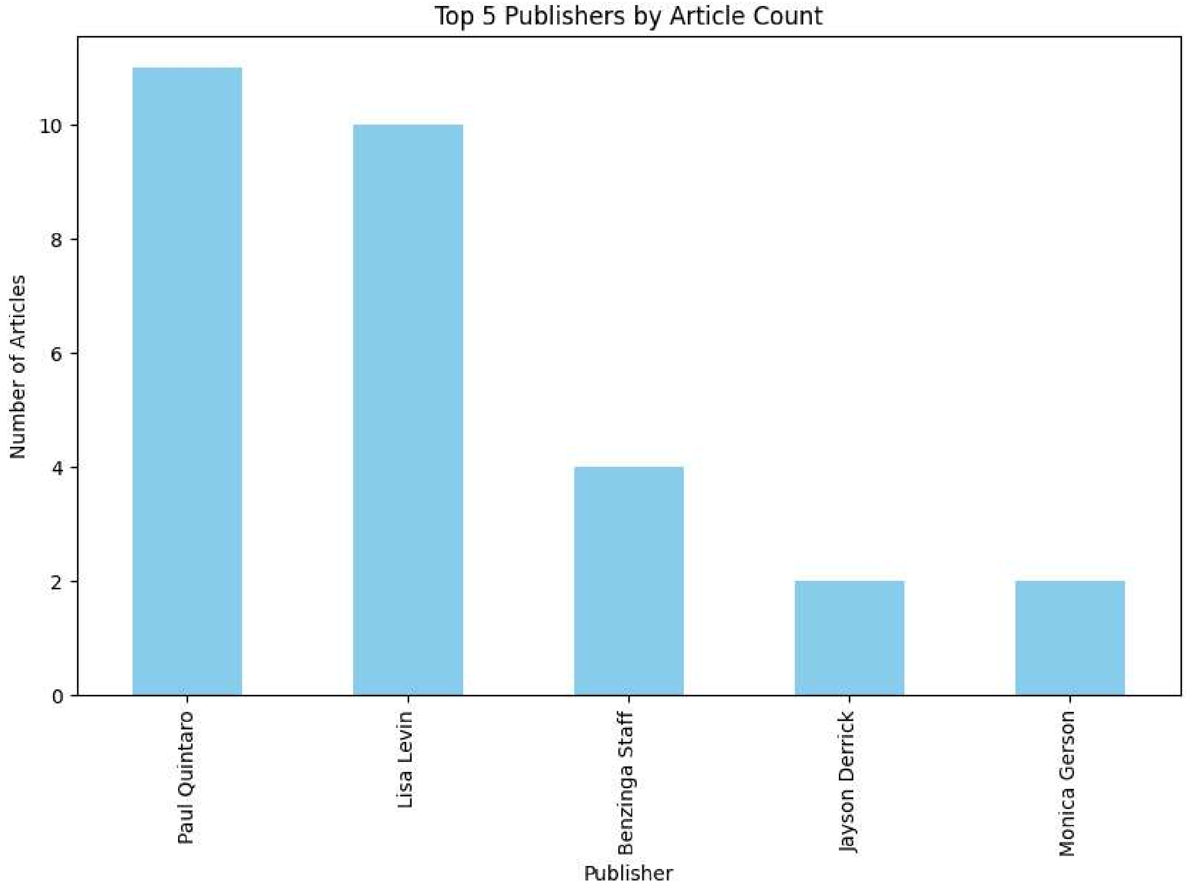


```
Top 5 words: {'in': 12, 'for': 11, 'to': 11, 'on': 8, 'stocks': 8}
```

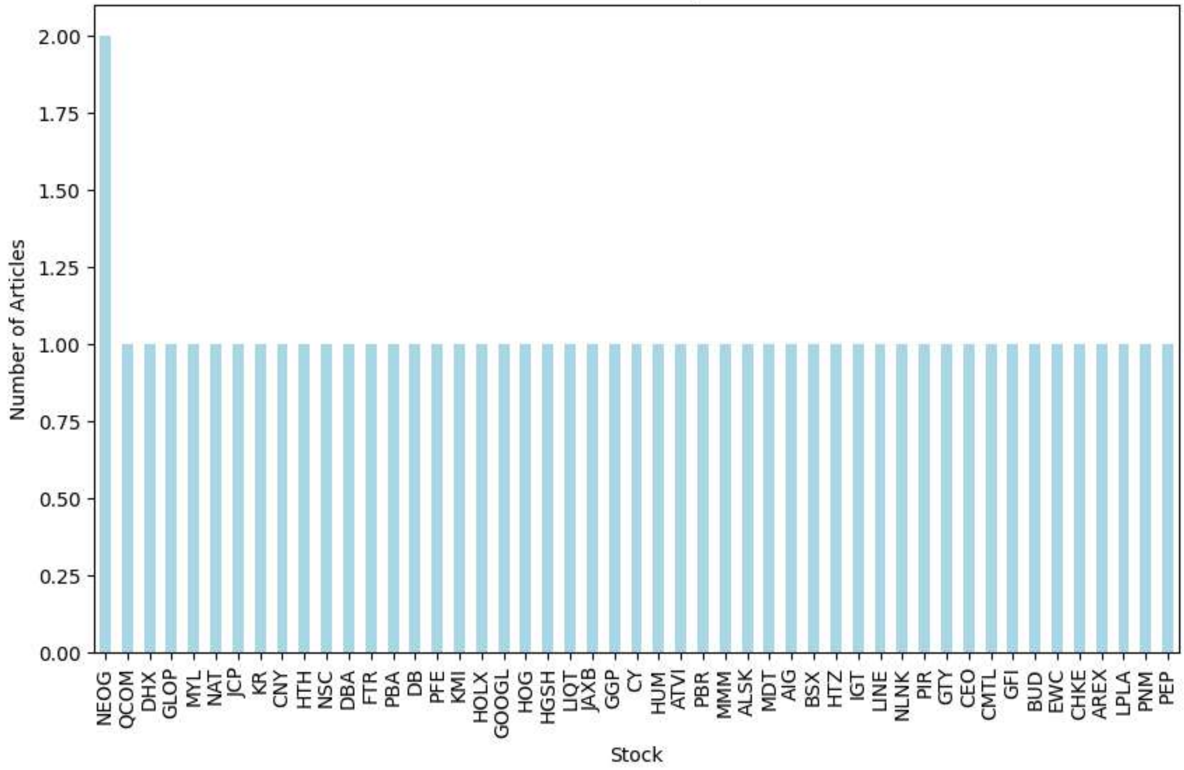
Top 5 Frequent Keywords in Headlines

for
in
stocks
on
to


Publication Frequency Over Time
Number of Articles
2.0
1.8
1.6
1.4
1.2
1.0
Date
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020


Top 5 Publishers by Article Count
Number of Articles
10
8
6
4
2
0
Paul Quintaro
Lisa Levin
Benzinga Staff
Jayson Derrick
Monica Gerson
Publisher

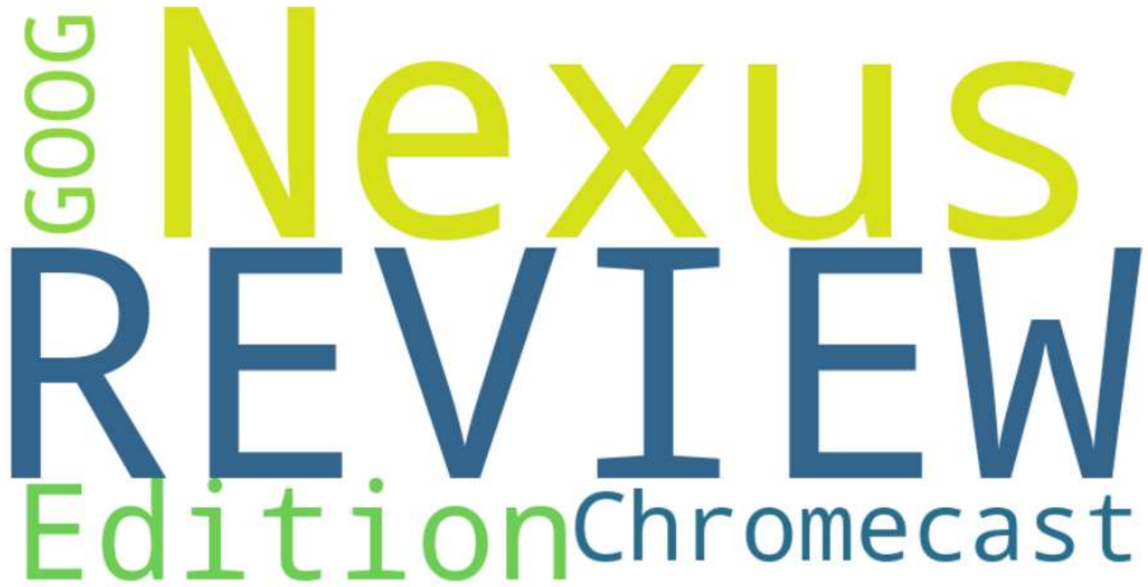
Number of Articles per Stock



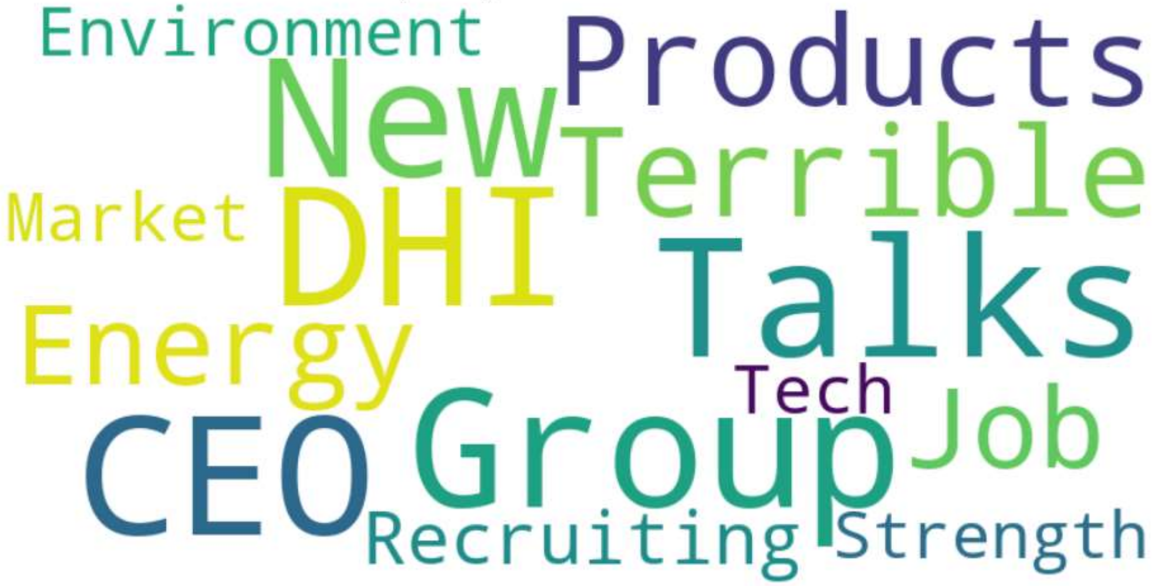
Frequent Keywords in Headlines for NEOG

Earnings
January
Scheduled
September

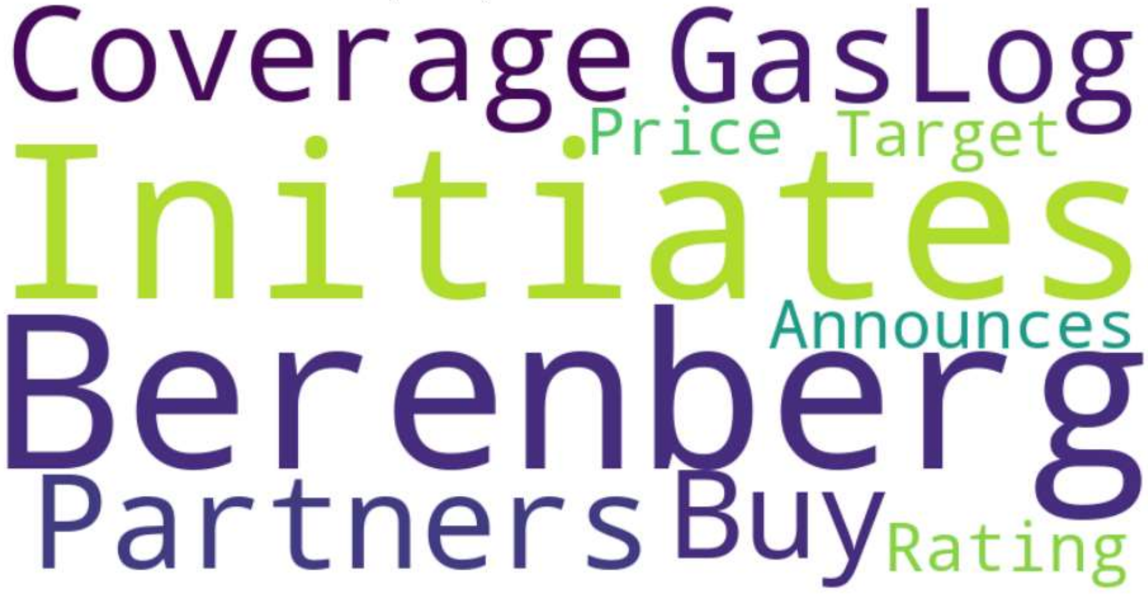
Frequent Keywords in Headlines for QCOM



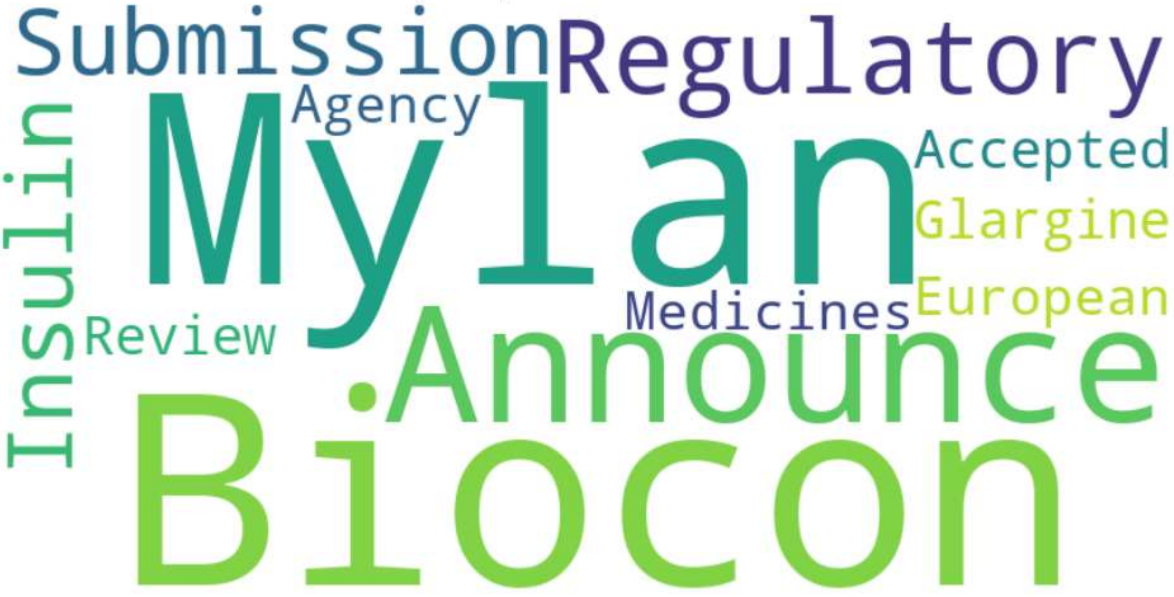
Frequent Keywords in Headlines for DHX



Frequent Keywords in Headlines for GLOP



Frequent Keywords in Headlines for MYL


Submission
Regulatory
Mylan
Agency
Insulin
Accepted
Glargine
European
Medicines
Review
Announce
Biocon