# Executive Summary

The media can have quite an influence on the public’s opinion about certain topics. That is no different when considering investor attitudes towards company stock market performance. The purpose of this paper is to examine how authors’ opinions and attitudes portrayed through news articles affect the performance of stocks within the S&P 500 index. The analysis focuses on the various readability index scores, word counts and structures, market cap sizes, and finally sentiment scores from a day’s news to predict the stock price movement of the following day.

The main challenge of this project was scaling computations from a few months to an entire year of data. This task occurred throughout the duration of the project from loading data in from GDELT, to quantifying sentiment and readability scores. To resolve this issue, I turned to AWS EC2 cloud computing environment.

**Add final results and findings…**

The next steps to this project would be to broaden the scope of the project in regard to news sources, expand the timeframe of historical data and news contents, and also vary the prediction horizon. In this project there were only three websites used to collect news articles from but increasing this number would allow for more opinions on news topics. Another way to develop this project in the future would be to collect additional years of both news and stock price data beyond 2021. Finally, investigating different prediction timelines would be another strategy for predicting stock prices. An example of this would be to use a week’s or month’s worth of data to predict the next week’s or month’s stock price movement. This could be more beneficial for long-term investor rather than a short-term investor.

# Introduction

# Data

This project required several types of data to be collected from various sources. The criteria that data covered was the news articles, S&P 500 index composition, stock open and close prices, and stock market cap values.

The main dataset that includes the news article details was collected from the GDELT 2.0 Event Database. The data spans the entire year of 2021 from January 1st to December 31st. Before processing the GDELT data contained 26 columns including columns regarding dataset ID, website, URL, title, people, topics, etc. After processing I retained three columns: dataset ID which I converted to a date stamp, website, and URL. The date column is used to specify input and output rows when trying to use a day’s news to predict the next day’s stock price movement. The website column is used to filter specific websites along with being an input parameter for a web scraping template to collect news article contents. Finally, the URL column is also used in the web scraping process as a direct link to the web page that is being scraped.

The S&P 500 index stock composition was pulled from Wikipedia. The stock open and close prices, along with the market cap values were gathered from Yahoo Finance. Although only S&P 500 companies are included in this analysis indicating a focus on large companies, an additional column was added to the market cap size of the companies specifying big, bigger, or biggest which is broken down by market cap values less than $50 billion, $50-100 billion, and over $100 billion. **Specify any additional price calculations such as movement**

Some issues that arised during the data collection process included the pure volume of data and inconsistent formats.

# Methodology

Web scrape news

Sentiment

Readability scores/Word counts

AWS EC2

**Add modeling**

# Results

# Conclusion