Twitter Analysis: Project 2 Presentation

Ryan Downing, Stefan Obradovic, Sahil Goel, Candace Austin, Rohan Samy, Kafana Ouattara

Dataset Description

We collected tweets from individuals on Twitter regarding three separate industries in the economy by grabbing tweets with explicit ticker mentions

- Twitter posts made between March 01, 2020 and March 12, 2022
- At most 10 tweets per analyst

Technology:

- AAPL, MSFT, GOOG, AMZN, TSLA, FB, NVDA, AVGO, CSCO, ADBE, CRM, ORCL, ITC, AMD, NFLX, TXN, INTU, PYPL, NOW, IBM, QQQ, TQQQ, ARKK

Financials:

- V, MA, AXP, PYPL, JPM, BAC, WFC, C, COF, PNC, USB, BRK-B, BRK-A, SCHW, MS, GS, MET, PRU, BK, TROW

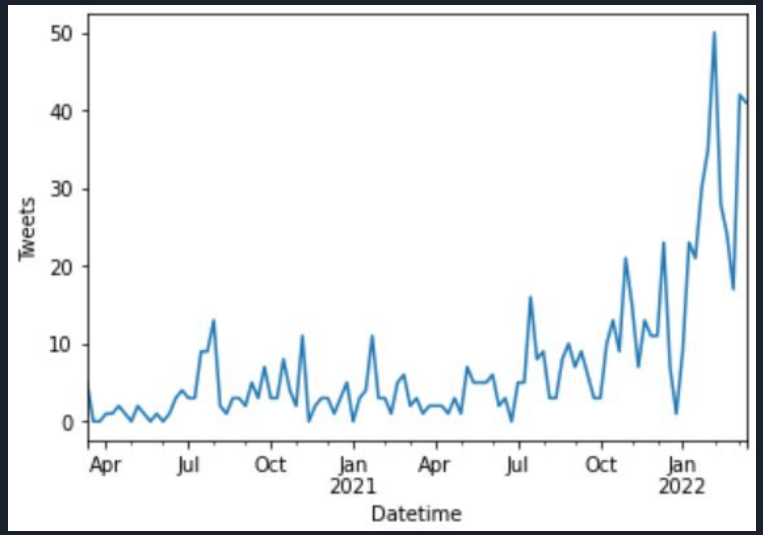
Energy:

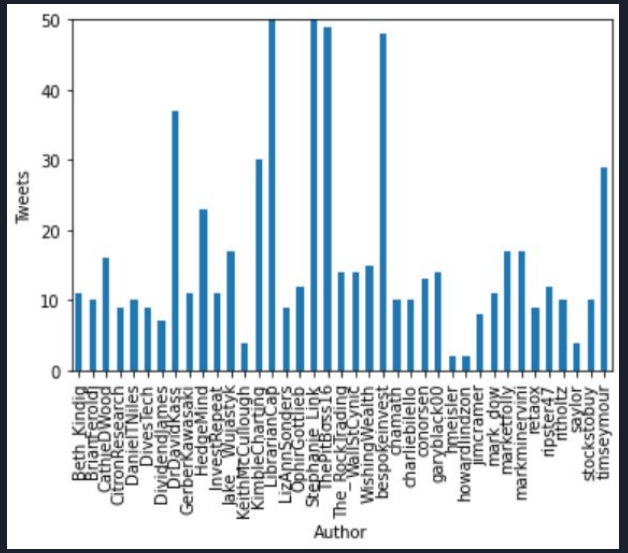
- XOM, CVX, SHEL, PTR, TTE, COP, EQNR, BP, PBR, ENB, CNQ, SNP, EOG, SLB, PXD, EPD, OXY, TRP, E, MPC, SU, KMI, VLO, WMB, ICLN, SPWR, ENPH, CWEN, SHFT, AMRC, XLE

Collecting the Dataset

Used Python's tweepy Twitter API

- Collected list of twitter usernames from active users in the financial analyst community.
- Obtained all tweets which contained relevant cashtags (i.e. $V, $MA).
- Subsampled twitter users who had many tweets for later analysis to remove bias.





Punctuation, Numbers, and Syntax

Cleaning text before applying analysis

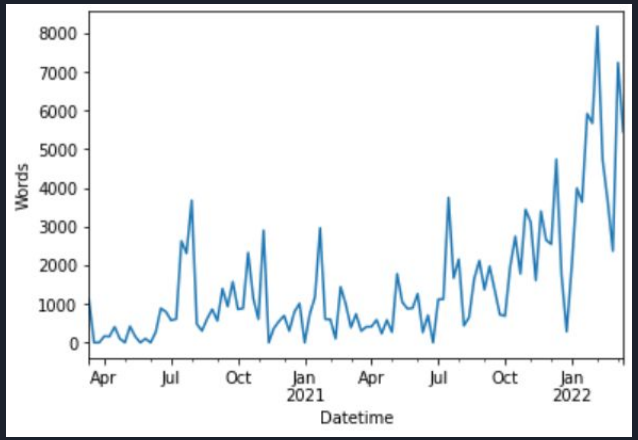
- Removed all characters which are not alphabetical or whitespace (including emojis and cashtags).
- Use NLTK's porter stemmer to get the roots of all words.
- Split on whitespace to get tokens, remove stopwords, and recreate sentence.

```
stemmer = PorterStemmer()
sw = set(stopwords.words("english"))

def stem_and_remove_sw(tokens):
  return [stemmer.stem(token) for token in tokens if stemmer.stem(token) not in sw]

data["cleaned_text"] = data["Text"].str.replace("[^A-Za-z\s]", "", regex=True).str.lower()

data['tokens'] = data["cleaned_text"].str.split(',')
data['tokens'] = data['tokens'].apply(stem_and_remove_sw)
data["tokenized_sentence"] = data["tokens"].apply(lambda tokens: ' '.join(tokens))
```

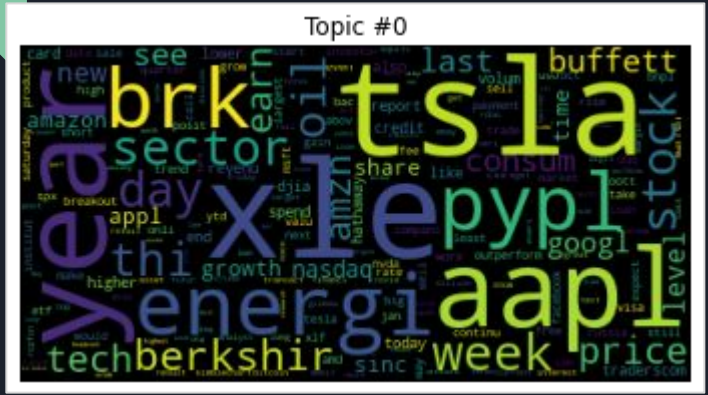


Gensim

```
import pandas as pd
import re
from nltk.stem.porter import *
from nltk.corpus import stopwords
from pandas.tseries import offsets
from wordcloud import WordCloud
from gensim.corpora.dictionary import Dictionary
from gensim.models.ldamodel import LdaModel
from pathlib import Path
import matplotlib.pyplot as plt
%matplotlib inline
```

	Author	Text	Date	Stock(s) Mentioned	Tweet Link	Misc.	Topic
0	Stephanie_Link	Didn't do the wage increase situation justice....	2021-05-24 19:34:08+00:00	FB	https://twitter.com/twitter/statuses/139691241...	NaN	Technology
1	bespokeinvest	Apple $AAPL now up 14% over the last month.	2021-07-14 14:35:44+00:00	AAPL	https://twitter.com/twitter/statuses/141531910...	NaN	Technology
2	HedgeMind	$AMD Unstoppable momentum, up 11%+ after repor...	2021-11-08 19:49:26+00:00	AMD	https://twitter.com/twitter/statuses/145779742...	NaN	Technology
3	jimcramer	critical reversal day and then $AMD announces ...	2022-02-24 21:22:06+00:00	AMD	https://twitter.com/twitter/statuses/149695863...	NaN	Technology
4	Beth_Kindig	Today we recap this week's biggest tech earnin...	2021-05-21 14:25:35+00:00	CSCO	https://twitter.com/twitter/statuses/139574760...	NaN	Technology

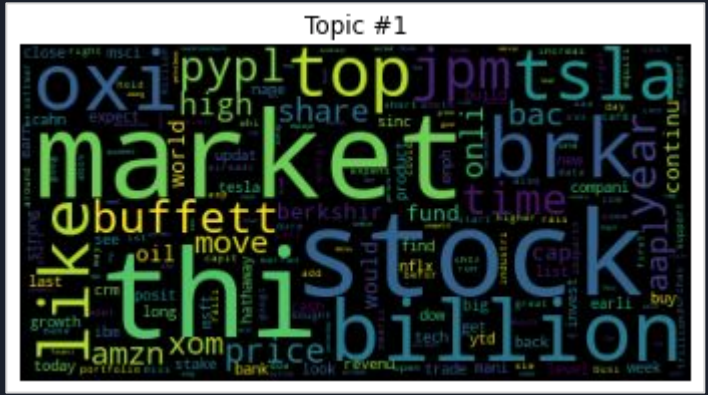
Word Cloud



Topic 0: Energy Stocks

Topic 1: Tech Stocks

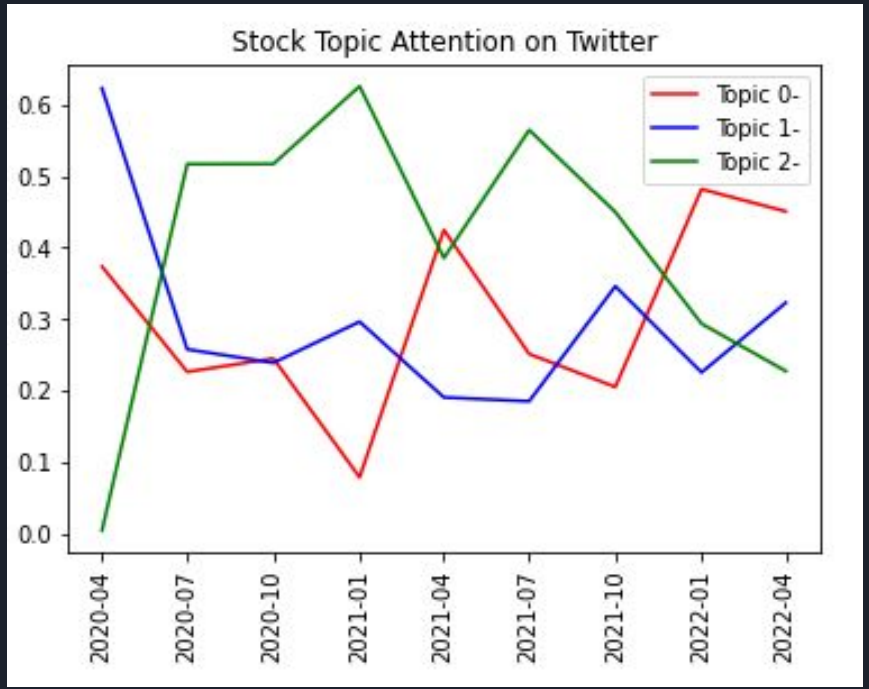
Topic 2: Financial Stocks



Document Topic Matrix

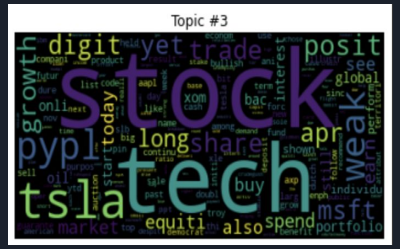
	0	1	2
0	(0, 0.37323412)	(1, 0.62279147)	(2, 0.0039743935)
1	(0, 0.22588567)	(1, 0.25737366)	(2, 0.5167407)
2	(0, 0.24457252)	(1, 0.23853293)	(2, 0.5168945)
3	(0, 0.07829677)	(1, 0.2960779)	(2, 0.6256254)
4	(0, 0.42456445)	(1, 0.19016759)	(2, 0.38526794)
5	(0, 0.25088882)	(1, 0.1846899)	(2, 0.5644213)
6	(0, 0.20474742)	(1, 0.3456513)	(2, 0.44960126)
7	(0, 0.48157552)	(1, 0.22509378)	(2, 0.29333073)
8	(0, 0.45047432)	(1, 0.32261384)	(2, 0.22691186)

Topics Over Time



Topic Model with K>3

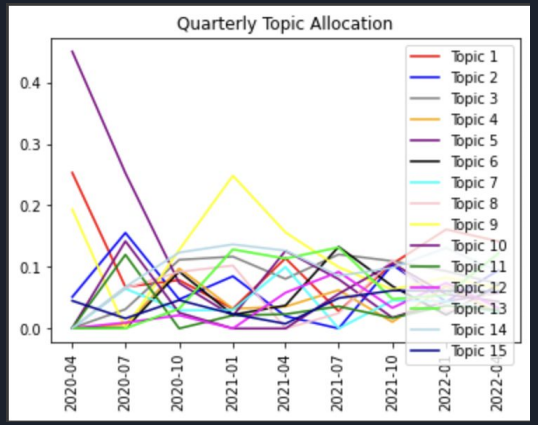
Topic Model was re-ran with K=15



Noticeable Changes:

1. Topics are more defined and easier to interpret
2. Topics include subcategories of the original three topics

	0	1	2	3	4	5	6	7	8	9	10	11
0	(0, 0.25312266)	(1, 0.051194035)	(2, 0.0007017556)	(3, 0.0007017555)	(4, 0.44999138)	(5, 0.0007017555)	(6, 0.0007017552)	(7, 0.00070175505)	(8, 0.19355984)	(9, 0.0007017555)	(10, 0.00070175546)	(11, 0.0007017557)
1	(0, 0.067472346)	(1, 0.15555872)	(2, 0.031350814)	(3, 0.0062164506)	(4, 0.25248697)	(5, 0.00017406476)	(6, 0.06471941)	(7, 0.065848425)	(8, 0.00017406506)	(9, 0.14190483)	(10, 0.11997226)	(11, 0.009573906)
2	(0, 0.07863966)	(1, 0.047320075)	(2, 0.11169524)	(3, 0.097573176)	(4, 0.07396668)	(5, 0.09256635)	(6, 0.029345956)	(7, 0.09200091)	(8, 0.12778364)	(9, 0.025668165)	(10, 4.3094264e-05)	(11, 0.02153308)
3	(0, 0.032087065)	(1, 0.08498815)	(2, 0.11679665)	(3, 0.032220133)	(4, 0.020444611)	(5, 0.022376962)	(6, 0.030284075)	(7, 0.102393925)	(8, 0.24826749)	(9, 6.150078e-05)	(10, 0.021413352)	(11, 6.1500774e-05)
4	(0, 0.114177644)	(1, 0.020048523)	(2, 0.08066652)	(3, 0.035580922)	(4, 0.12620741)	(5, 0.037426714)	(6, 0.09987485)	(7, 6.761345e-05)	(8, 0.15627933)	(9, 6.761339e-05)	(10, 0.023276767)	(11, 0.058228575)
5	(0, 0.028127292)	(1, 8.061291e-05)	(2, 0.12039878)	(3, 0.062453017)	(4, 0.08018695)	(5, 0.13280597)	(6, 8.0612874e-05)	(7, 0.024892371)	(8, 0.09963214)	(9, 0.05635226)	(10, 0.03600321)	(11, 0.092019655)
6	(0, 0.10498282)	(1, 0.10366948)	(2, 0.109607466)	(3, 0.01061175)	(4, 0.017761258)	(5, 0.06877794)	(6, 0.045513757)	(7, 0.11130331)	(8, 0.06314738)	(9, 0.10558411)	(10, 0.017221682)	(11, 0.03352547)
7	(0, 0.16099592)	(1, 0.043507446)	(2, 0.09333468)	(3, 0.06926284)	(4, 0.04394446)	(5, 0.022264361)	(6, 0.04426789)	(7, 0.026281612)	(8, 0.08283086)	(9, 0.05572702)	(10, 0.037132055)	(11, 0.07416638)
8	(0, 0.14157699)	(1, 0.09479937)	(2, 0.0725734)	(3, 0.024733096)	(4, 0.05825246)	(5, 0.071313135)	(6, 0.02617128)	(7, 0.06285469)	(8, 0.06571887)	(9, 0.043677084)	(10, 0.029515153)	(11, 0.033200055)



Topic Model Prediction

Predictions of Which Topic Should have the Highest Weight

Tweet 1: Technology
Tweet 2: Finance
Tweet 3: Finance
Tweet 4: Energy
Tweet 5: Finance

Topic Model Weight Predictions

	Unnamed: 0	User	Tweet	clean_tokens	0	1	2
0	0	ARKInvest	ARK's updated open-source Tesla model yields a...	[ark, updat, open, sourc, tesla, model, yield,...	0.025014	0.612669	0.362318
1	1	xxxx_chen	$PYPL bears getting louder \n\nhttps://t.co/0u...	[pypl, bear, get, louder, uhxdvcdau, thi, doe,...	0.016240	0.071141	0.912618
2	2	bishnuvardhan	$PYPL waking up and outperforming $QQQ. Poten...	[pypl, wake, outperform, qqq, potenti, someth,...	0.413838	0.266361	0.319800
3	3	clpirtle25	$XOM NEW ARTICLE : Exxon sees carbon capture m...	[xom, new, articl, exxon, see, carbon, captur,...	0.090699	0.635759	0.273543
4	4	MYFORTPIERCE	I had someone tell me they are convinced that...	[someon, tell, convinc, run, bank, point, info...	0.281458	0.318613	0.399929

Topic Model's Highest Weighted Topic

Tweet 1: Technology
Tweet 2: Finance
Tweet 3: Energy
Tweet 4: Energy
Tweet 5: Finance

Paper Summary - Background

- Climate change presents risks to investors that cannot be easily hedged with futures or insurance contracts due to the widespread implications of climate disasters

- The authors propose an approach to hedge the risk of climate change in the long term by utilizing text mining techniques on climate change news

Paper Summary - High Level Strategy Overview

- Main idea is to construct a portfolio of publicly traded U.S. equities that hedges short-term climate change news in order to be protected against the dynamic nature of climate change in the long run
- Strategy follows a dynamic hedging approach created by Black, Scholes and Merton
- Two indices are created which measure the level and sentiment of climate related news based on the Wall Street Journal and data analytics vendor Crimson Hexagon (compilation of >1000 news outlets)
- From these indices, the portfolio is constructed by detecting stocks with returns that are negatively correlated with the level of negative news pertaining to climate
- This portfolio is frequently updated to account for recent news, thus protecting the investor overtime

Paper Summary - Climate Change Risk Indexes

- Climate Change Vocabulary (CCV) Corpus
  - 19 Climate change White-papers from government agencies and conferences
  - 55 climate change word glossaries
  - Constructed by counting the frequency of stemmed words in the corpus (tf-idf matrix calculated using corpus as a single document)
- WSJ Index
  - Constructed by computing cosine similarity between tf-idf of CCV corpus and Wall Street Journal Articles
  - No separation between positive and negative sentiment regarding climate change news

Figure 1: Climate change vocabulary

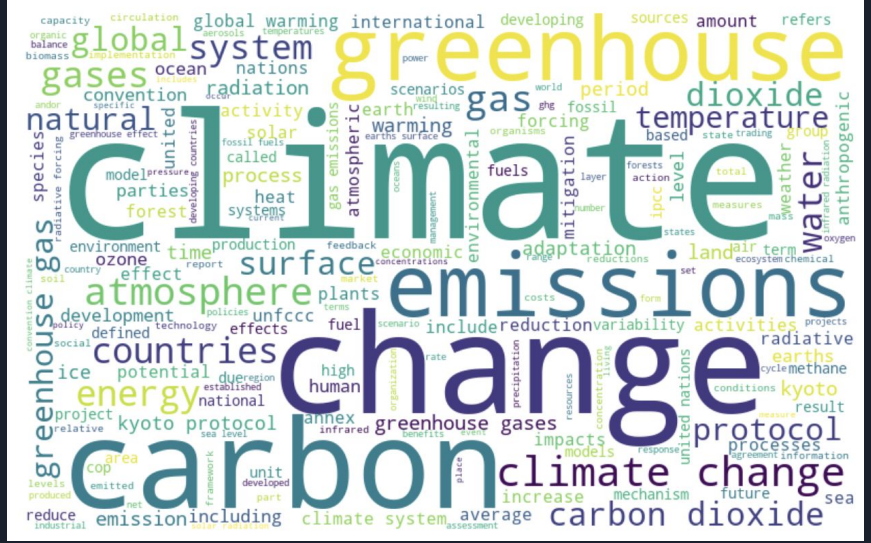
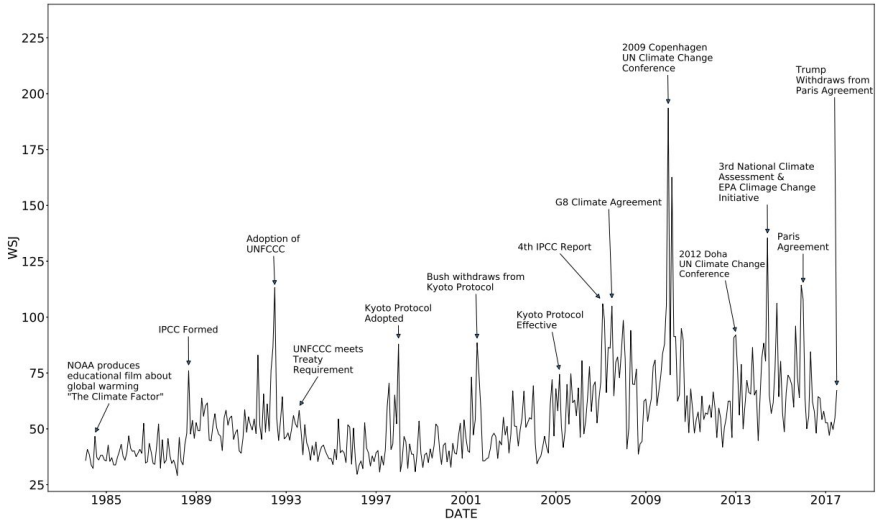


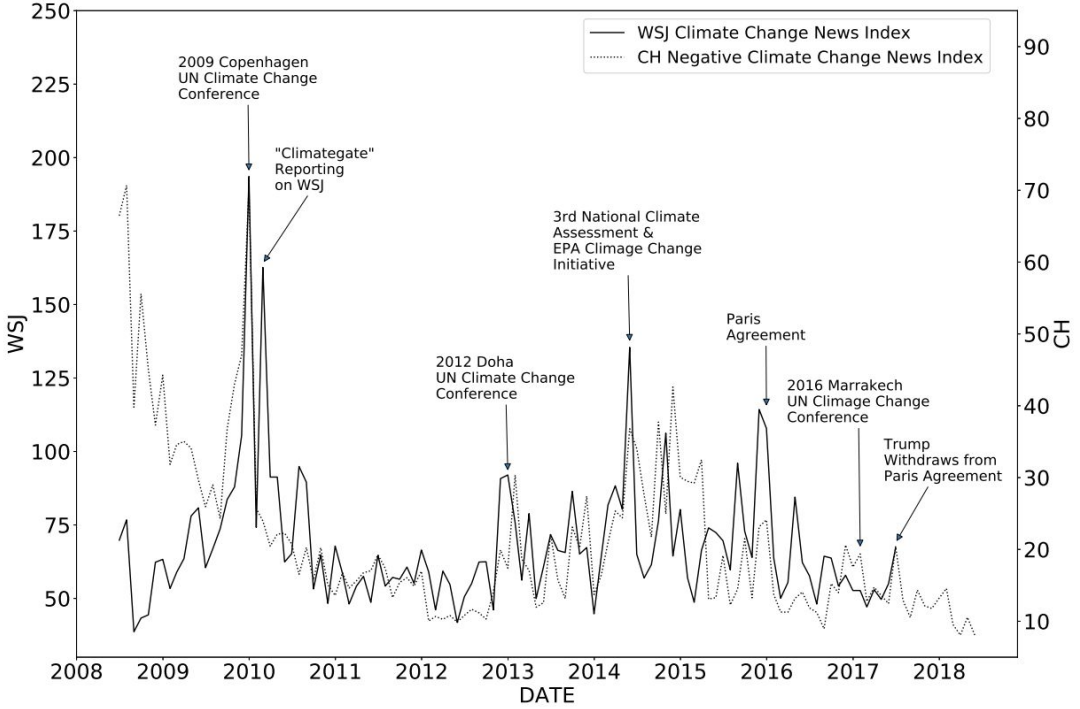
Figure 2: WSJ Climate Change News Index



Paper Summary - Climate Change Risk Indexes

- Crimson Hexagon (CH) Negative Climate Change News Index
  - Constructed using proprietary tool that calculates sentiment regarding a particular topic across many news outlets
  - Computed magnitude of negative sentiment regarding “climate change”
- Hedging Portfolios
  - Long companies that perform well when climate change news is up
  - Short companies that perform poorly when climate change news is up
  - Constructed using correlation between company ESG metrics
  - Reduce exposure to other risks to allow for climate change hedging that does not affect other market risks.

Figure 3: CH Negative Climate Change News Index



Paper Summary - Conclusions

- While the authors argue that their hedging approach is successful, they recognize that their strategy isn't definitively the "best" and that their findings should be used as a starting point for further research

- More research should be done to account for the fact that there are different types of climate risk. For example, this methodology does not distinguish between news pertaining to physical climate risk and regulatory climate risk

- Additionally, more types of assets could be considered in the hedging portfolio such as ETFs and international equities

Thank You!!