2 - Large Language Models and Transformers

Giovanni Della Lunga
giovanni.dellalunga@unibo.it

Halloween Conference 2024

Bologna - October, 30 2024

NLP in Emerging Risk Analysis

- In this part we are going to see how Natural Language Processing can be used to address the analysis of various dimensions of emerging risk, focusing on climate, geopolitical, and cybersecurity risks
- In particular we present 5 papers in which the previously described techniques are applied to build indexes and to study the relationship with market data ...

NLP in Emerging Risk Analysis

Climate Risk (Kolbel et al., Engle et al.):

- These presentations explore climate risk metrics, discussing the use of machine learning to evaluate climate-related financial exposures.

Geopolitical Risk (Caldara et al.):

- Caldara's analysis delves into the Geopolitical Risk Index (GPR), illustrating its construction through a text-based approach that captures global events' impacts on markets.

Cybersecurity Risk (Florakis et al., Jamilov et al.):

- The cybersecurity presentations introduce novel measures for firm-level cyber risk, applying computational linguistics to financial disclosures and earnings calls. They showcase how textual analysis can uncover cybersecurity risk exposures, assess market implications, and help investors factor these into decision-making.

Subsection 1

BERT and Climate Risk Disclosure Impact on CDS Term Structure

J. F. Kolbel, M. Leippold, J. Rillaerts, Q. Wang

Introduction to the Research Goal

- The study aims to quantify the **impact of regulatory disclosures of transition and physical climate risks on the credit default swap (CDS) market** using the BERT model.
- Climate risks entail physical risks that emerge from extreme weather events and transition risks stemming from regulatory reforms intended to combat global warming. These two risks may affect companies in different ways.
- The directional effect of disclosing climate risks on risk premia is not obvious. Risk disclosure may trigger two opposing effects, a risk-perception and an information-uncertainty effect.

Introduction to the Research Goal

Key points:

- Use of an advanced LLM like BERT for classification;
- Firm level analysis (text analysis of 10-K reports)
- Fine tuning of pre-trained LLM
- Analysis of impact on market data (CDS)

Data Collection and Processing

10-K Report

What Is Form 10-K?

Form 10-K is a comprehensive report filed annually by a publicly traded company about its financial performance and is required by the U.S. Securities and Exchange Commission (SEC). Some of the information a company is required to document in the 10-K includes its history, organizational structure, financial statements, earnings per share, subsidiaries, executive compensation, and any other relevant data.

KEY TAKEAWAYS

- A 10-K is a comprehensive report filed annually by public companies about their financial performance.
- The report is required by the U.S. Securities and Exchange Commission (SEC) and is far more detailed than the annual report.
- Information in the 10-K includes corporate history, financial statements, earnings per share, and any other relevant data.
- The 10-K is a useful tool for investors to make important decisions about their investments.

Data Collection and Processing

Data Sources:

- Regulatory Disclosures analyzed through 10-K filings.
- Market Data: CDS spreads to analyze market impact.

Data Processing:

- Application of NLP to quantify risk disclosures.
- BERT model applied to classify risk-related statements.
- BERT is used to classify sentences from the section Item 1.A of a firm's 10-K report and generate a firm-specific measure of both transition and physical risks.

Methodology

BERT for NLP:

- Differentiates sentences related to transition and physical climate risks.
- Deriving a measure of climate risk from these regulatory filings is a subtle task. Companies are flexible in how they disclose and describe their climate risk exposure, which could make it difficult to compare the filings relative to each other.

Index Construction: Developed indices for physical and transition risks based on term frequency and context.

Fine Tuning of BERT

- Pre-trained BERT fine-tuned with specific examples from financial disclosures.
- **Added layers** tailored for financial text classification.

Fine Tuning of BERT

BERT was fine-tuned specifically for sentence classification of climate risk disclosures, distinguishing between:

- Transition Risk: Risks arising from regulatory or policy changes related to climate action (e.g., new carbon tax, regulatory compliance).
- Physical Risk: Risks arising from the physical impacts of climate change (e.g., natural disasters, extreme weather events).
- General (non-climate related): Sentences not related to either transition or physical climate risks.

The classification task was structured as a multi-class classification problem where each sentence in a 10-K filing needed to be assigned to one of the three categories.

Fine Tuning of BERT

Source of Labeled Data:

- The training data was derived from sample reports provided by TCFD (Task Force on Climate-Related Financial Disclosures) guidelines, offering nearly 1,000 example sentences.
- In addition, the authors manually added random sentences not related to climate risks to ensure proper representation of general disclosures.

Manual Annotation of Sentences:

- Sentences were labeled manually by experts with knowledge of both finance and the TCFD guidelines. This manual step ensured high-quality training data with accurate labeling of risks as transition, physical, or general.

Fine Tuning of BERT

Confusing Sentence Review:

- The authors employed an iterative method to enhance the quality of the training data.
- The fine-tuned model was run on the 10-K dataset, and confusing sentences (where probabilities across categories were similar) were identified.
- These confusing sentences were reviewed by human annotators for proper classification, and added to the training set along with confident sentences to maintain the balance of the dataset.

Final Training Set Size:

- After several rounds of this review and annotation process, the authors built a final training set consisting of **3,192** classified sentences.

Fine Tuning of BERT

Network Architecture:

- The authors added two fully connected layers on top of the [CLS] token's output (which represents the sentence-level embedding).
- Both of these layers had 128 hidden units.

Sentence Classification and Index Calculation

Sentence Classification:

- After fine-tuning, the model was applied to classify sentences from 10-K reports (Item 1.A) into transition, physical, or general categories.
- For each sentence, BERT calculated a probability distribution across the three classes.
- A threshold of 0.8 was used to determine whether a sentence was confidently assigned to one of the climate risk categories (transition or physical).
- Sentences with lower confidence remained unclassified or were assigned to the "general" category.

Sentence Classification and Index Calculation

Document-Level Aggregation:

- The final climate risk scores for each company were generated by averaging the binary classification results of individual sentences in each company's 10-K filing.
- This yielded a document-level risk score that reflected the relative importance of transition and physical risks compared to other disclosed risks.
- The papers do not appear to define "transition risk score" and "physical risk score" explicitly. The described methodology seems to imply that these scores are derived by aggregating BERT classifications on sentences related to either transition or physical risks in the 10-K filings.

Sentence Classification and Index Calculation

- For example, the scores can be calculated as the proportion of sentences in each category relative to the total, representing the company's focus on each risk type in its disclosures.

Linking Climate Risk Indices to CDS Market Data

CDS (Credit Default Swap) Spreads:

- CDS spreads represent the cost of insuring a firm‘s debt against default.
- A higher spread indicates a higher perceived risk of default.
- Thus, CDS spreads are sensitive to the market‘s perception of a firm”s risk profile.

Data Source for CDS:

- The CDS data were collected from a reliable financial data provider, typically Thomson Reuters Datastream, covering various firms across time with daily observations.

Linking Climate Risk Indices to CDS Market Data

Regression Analysis:

- The authors employed classical statistical regression models to analyze the relationship between the climate risk indices and changes in CDS spreads
- By including the transition and physical risk indices as explanatory variables, they could observe whether higher scores in either index corresponded to higher (or lower) CDS spreads.

Linking Climate Risk Indices to CDS Market Data

They analyze how the regulatory disclosure of transition and physical risks impacts CDS spreads over various maturities. Specifically, they estimate the following one-month ahead forecasting regressions:

$$\Delta S^m_{i,t+1} = \beta_T \Delta CR\text{-}Transition_{i,t} + \beta_P \Delta CR\text{-}Physical_{i,t} + \Phi \Delta X_{i,t} + \Theta \Delta Y_t + \epsilon_{i,t+1}$$

where

- $S^m_{i,t+1}$ denote the next month's (average) m-year spread
- $X_{i,t}$ and Y_t are firm-specific and macro-economic control vectors, respectively
- $CR\text{-}Trans_{i,t}$ and $CR\text{-}Phys_{i,t}$, denote BERT-based proxies for transition and physical risk.

Insights on Market Pricing of Transition and Physical Risks

The results of the regression analysis provided key insights into how the market prices climate risks:

- Transition Risk: An **increase in transition** risk disclosures was generally associated with **higher CDS spreads**, especially after landmark policy events like the Paris Agreement (2015). This reflects a market perception that **regulatory changes may impose additional compliance costs or operational risks, thus increasing the firm's credit risk**.
- Physical Risk: **Higher physical** risk disclosures were often associated with **lower CDS spreads**. The authors suggest this might be because **physical risks are perceived as more predictable and manageable (relative to regulatory risks)**, which could reduce uncertainty for investors and hence lower perceived credit risk.

Applications to Asset Pricing

Potential for Broader Application:

- The link between climate disclosures and CDS spreads can be used to refine **asset pricing models**, allowing analysts and investors to incorporate climate risk into credit risk assessments and pricing.

Climate Risk as a Factor in Investment Decisions:

- The study's findings suggest that climate risk disclosures-particularly transition risks-are material to credit risk and, by extension, asset pricing. This could lead to more systematic inclusion of climate risk factors in investment models and decision-making.

Summary

- In essence, this approach offers a novel methodology to translate qualitative climate risk disclosures into quantitative indices that reflect transition and physical risks.
- By linking these indices to CDS spreads, the authors provide a framework to understand how climate risks are priced in the credit markets, revealing that the financial impacts of regulatory and physical risks are distinct and can influence investor behavior differently.
- This work underscores the growing importance of climate risk as a factor in financial markets, particularly in asset pricing and credit risk assessment.

Subsection 2

Hedging climate change news

R. F. Engle, S. Giglio, H. Lee, B. Kelly, J. Stroebel

Main Goal of the Research

Purpose

- Propose a method to dynamically hedge climate change risk using portfolios.
- Extract climate change news from textual analysis.
- Hedge climate risk based on news information.

Introduction to the Research Goal

Key points:

- Global level analysis
- Use of News Data (mainly Wall Street Journal)
- Analysis of impact on market data

Example:

If Company A's 10-K filing contains 100 sentences, and 25 sentences discuss transition risks, the **transition risk score** for Company A will be:

$$Transition\ Risk\ Score = \frac{25}{100} = 0.25$$

This means that 25% of the company's 10-K filing discusses transition risks. Similarly, if 15 sentences are about physical risks, the **physical risk score** would be:

$$Physical\ Risk\ Score = \frac{15}{100} = 0.15$$

In the end, these scores represent how much the company is disclosing about climate risks, with a higher score indicating a greater focus on such risks.

Problem Addressed

Challenge:

- Long-term, non-diversifiable nature of climate risk.
- Lack of traditional insurance mechanisms for climate disaster events.

Solution Proposed:

- Hedging via Portfolios
- Construct portfolios to hedge climate change news.
- Use equities to build dynamic hedge portfolios.

Climate News and Risk

News-Based Approach
Two indices developed

- WSJ Climate Change News Index: Measures climate change news intensity in The Wall Street Journal.
- CH Negative Climate Change News Index: Focuses on negative climate change news sentiment from a broader news dataset.

Data Sources:

- Textual data from The Wall Street Journal (WSJ) and Crimson Hexagon.
- Stock market data from CRSP for U.S. equities.

Methodology Overview

1 - Climate News Series Construction:

- Use textual analysis on news articles to create a time series reflecting climate risk updates.
- The two indices capture climate-related news intensity and sentiment to reflect innovations in climate risk.

2 - ESG Scores as Proxies:

- MSCI and Sustainalytics Scores: ESG scores serve as proxies for firms' exposure to climate risk.
- Portfolio construction relies on firm-level characteristics like carbon footprint and environmental performance.

3 - Mimicking Portfolio Approach:

- The climate news factor is projected onto stock returns, forming hedge portfolios that correlate with climate news innovations.

Climate News Indices Development

WSJ Climate Change News Index:

- Based on cosine similarity between WSJ article content and a climate change vocabulary created from authoritative climate documents.
- Key climate events like the Paris Agreement and Copenhagen Conference show clear spikes.

CH Negative Climate Change News Index:

- Uses sentiment analysis to focus on negative climate change news from over 1,000 outlets.
- Aims to better capture investor concerns about worsening climate risk.

WSJ Climate Change News Index

- This index is designed to measure the frequency and intensity of climate-related topics in The Wall Street Journal (WSJ), which is widely read by the investment community.
- Let's see how it's constructed...

WSJ Climate Change News Index

A. Climate Change Vocabulary (CCV) Creation

- The authors collected 74 authoritative documents on climate change from institutions like IPCC, EPA and NASA.
- These texts were processed to create a Climate Change Vocabulary (CCV), which includes frequently occurring terms and phrases related to climate change, such as global warming, carbon emissions, sea level rise, etc.
- Terms were reduced to their root forms (using "stemming") to capture variations (e.g., "warming" and "warmed" both count as "warm").

WSJ Climate Change News Index

A. Climate Change Vocabulary (CCV) Creation

- Each term was then represented by term frequency-inverse document frequency (tf-idf), meaning terms that appear frequently across all documents have lower weights, while terms rare in other contexts have higher weights. This allows the vocabulary to reflect core climate change terminology.

WSJ Climate Change News Index

B. Matching WSJ Content to CCV

- Every daily WSJ edition was analyzed and converted into tf-idf scores, which quantified how closely each edition matched the CCV.
- Cosine Similarity: This method was used to compare each day's WSJ content with the CCV.
- Days where WSJ content is very similar to the CCV (i.e., heavily focused on climate change) score close to 1, while days with little to no climate-related content score close to 0.
- The WSJ Climate Change News Index is thus a scaled version of these cosine similarity scores, representing the proportion of WSJ's content related to climate change each day.

WSJ Climate Change News Index

C. Scaling and Smoothing

- The resulting index values are scaled by a factor of 10,000, making it easier to interpret changes over time.
- This daily index is then averaged to a monthly level to reduce noise and facilitate further analysis.

WSJ Climate Change News Index

CH Negative Climate Change News Index

This second index is specifically focused on negative sentiment climate news and uses data from a broader range of sources. It includes news articles across over 1,000 outlets like The New York Times, BBC, CNN, and Reuters. The process to construct this index was as follows:
A. Data Collection via Crimson Hexagon

- The authors collaborated with Crimson Hexagon (CH), a data analytics provider that archives a massive corpus of news and social media posts.
- The search term "climate change" was applied to CH's dataset to retrieve climate-related articles starting from 2008.
- To focus solely on reputable news media, only articles from major outlets were included, excluding social media posts.

CH Negative Climate Change News Index

B. Sentiment Analysis

- CH's proprietary sentiment analysis tools were used to categorize articles based on sentiment, particularly identifying those with a negative sentiment around climate change.
- This approach assumes that when there is more negative news on climate change, investors are likely to perceive an increase in climate risk.

CH Negative Climate Change News Index

C. Index Construction and Scaling

- The CH Negative Climate Change News Index was then calculated as the proportion of all news articles related to climate change that had a negative sentiment.
- Similar to the WSJ index, this measure was scaled by a factor of 10,000 to facilitate interpretation.

D. Comparing and Validating the Indices

- Both indices were observed to spike around major climate events like the Kyoto Protocol, Paris Agreement, and other global climate conferences.
- A correlation analysis was conducted between the two indices, which revealed a moderate correlation (about 0.3), indicating that both capture common elements of climate risk but are distinct in how they respond to different news events.

CH Negative Climate Change News Index

Why they need a Negative Sentiment Index

- The authors introduce a negative sentiment climate change news index to address potential limitations of the WSJ Climate Change News Index. Specifically, this negative sentiment index is necessary to focus on the bad news or increased risk associated with climate change, which can have unique implications for financial markets
- Climate change news is not uniformly negative. Some reports may discuss advancements in green technology or regulatory progress that mitigate risks, which can actually be perceived as positive for long-term investors.

Why they need a Negative Sentiment Index

- If all climate change news is treated equally, the hedge portfolio could overreact to any climate news, even if it reflects beneficial developments rather than heightened risks.
- The negative sentiment index isolates only the worsening climate news, helping investors hedge specifically against increases in perceived climate risks.

Summary of the Indices‘ Roles

- Together, the WSJ Climate Change News Index and the CH Negative Climate Change News Index offer two perspectives on climate news.
- The WSJ index captures general climate news intensity as reflected in a key financial publication.
- The CH index focuses on negative news sentiment from a broader media landscape.
- These indices thus serve as dynamic proxies for climate risk innovations, allowing the authors to test the effectiveness of different hedge portfolios based on the two indices‘ innovations.

Setting Up the Climate News Innovations (Hedge Targets)

The first goal is to create hedge targets-variables that capture unexpected changes (innovations) in climate news. They calculate innovations by detrending the indices using an autoregressive (AR(1)) model:

- Each monthly value of the WSJ and CH indices is regressed against its own prior month value.
- The residuals (differences between predicted and actual values) represent innovations in climate news for that month.
- These residuals, now de-trended, serve as the hedge targets denoted CC_t^{WSJ} and $CC_t^{NegNews}$ for the WSJ and CH indices, respectively.
- These targets reflect new information about climate risk that investors did not anticipate based on prior trends.

Selecting the Hedge Portfolio Assets

The authors then need to select assets that are likely to respond to these climate news innovations.

- U.S. Equities Focus: They use U.S. stock market equities since stocks are liquid, readily tradable, and have diverse industry and firm-level characteristics that could reflect varying exposures to climate risks.
- Screening Stocks: Only common equity stocks from the NYSE, AMEX, and NASDAQ exchanges are included, with filters applied to avoid microcaps and penny stocks (below \$5/share) to maintain liquidity and avoid market anomalies.

Selecting the Hedge Portfolio Assets

Defining Climate Exposure Characteristics Using ESG Scores:

- MSCI and Sustainalytics ESG Scores serve as proxies for each firm's climate risk sensitivity.
- These scores capture environmental impacts, such as carbon footprints or energy efficiency, that are expected to correlate with climate risk sensitivity.
- The ESG scores allow the authors to sort stocks based on their environmental performance, forming portfolios that include "green" (environmentally friendly) or "brown" (environmentally harmful) firms.

Mimicking Portfolio Construction

Factor Model Setup:
They start with a linear factor model:

$$r_t = \beta_{CC} \times CC_t + \beta \times v_t + u_t$$

where:

- r_t: Vector of excess returns of stocks.
- β_{CC}: Sensitivity of each asset to the climate news factor CC_t? β: Sensitivity to other market factors v_t? (e.g., size, value)
- u_t: Idiosyncratic error term.

This model is used to project the climate news factor onto stock returns by using the mimicking portfolio weights estimated through a cross-sectional regression.

Mimicking Portfolio Construction

Incorporating Standard Market Factors:

- To ensure that the hedge portfolios are only exposed to climate risk (and not other common market factors), they add size, value, and market factors in the mimicking portfolio regression.
- This approach ensures that the climate hedge portfolio is factor-neutral (uncorrelated with size, value, etc.) and captures only climate-related risks.

Mimicking Portfolio Construction

Estimating Mimicking Portfolio Weights:

- They regress each innovation in the climate news indices on the returns of these characteristic-sorted portfolios, creating portfolios that are long "winners" and short "losers" when climate news emerges.

Summary of Hedge Portfolio Characteristics

Finally, the authors analyze the characteristics of the hedge portfolios, revealing interesting findings:

- The portfolios are industry-balanced and do not simply reflect long positions in green energy or short positions in fossil fuels.
- Firms from diverse sectors show sensitivity to climate news, suggesting that climate exposure is not strictly industry-dependent.
- This construction enables the hedge portfolio to respond accurately to climate news, making it an adaptable financial instrument for managing climate-related financial risks.

Subsection 3

Measuring Geopolitical Risk

D. Caldara and M. Iacoviello

Main Goal of the Research

Purpose: Develop a news-based measure of geopolitical risk (GPR).

- They construct a newspaper-based indexes of geopolitical risk (GPR), daily and monthly, global and country-specific, and examine their evolution since 1900.
- Using aggregate macroeconomic data, then show that higher GPR increases the probability of an economic disaster and predicts lower investment and employment.
- Using firm-level data, they document that the adverse implications of geopolitical risk are stronger for firms in more exposed industries, and that high firm-level GPR is associated with lower firm-level investment.

Analyze how geopolitical events like wars and terrorism affect macroeconomic variables. Create an index to forecast economic downturns.

Definition of Geopolitical Risk

Geopolitical Risk (GPR):

- Defined as the threat, realization, or escalation of adverse geopolitical events (wars, terrorism, state tensions).
- Measured using news articles and public perception.

Motivation:

- Key determinant of investment, employment, and disaster risk.
- Historically linked with significant economic downturns, such as World Wars and 9/11.

Research Questions

Key Questions Addressed:

- How does geopolitical risk affect investment and employment?
- How can a systematic measure of geopolitical risk help forecast economic downturns?

Methodology Overview

Textual Analysis:

- News-based index created by analyzing articles from major newspapers.
- Dictionary-based method used to track mentions of geopolitical events.

Quantitative Models:

- VAR models to capture the impact of GPR on macroeconomic variables.

Data Sources

Newspaper Data:

- Collected from 25 million articles across 10 leading English-language newspapers.
- Historical data from 1900 (e.g., New York Times).
- Corresponding to about 30,000 and 10,000 articles per month in the recent and historical sample, respectively.

Other Data:

- Firm-level data from 135,000 earnings calls to track geopolitical risks.

Data Selection Criteria

Article Selection Process:

- Articles selected based on pre-defined terms related to war, terrorism, political conflict.
- Categories include words like 'nuclear,' 'conflict,' 'terrorism,' paired with 'threat' and 'invasion.'

Geopolitical Risk Index (GPR)

Index Construction:

- The recent GPR index starts in 1985 and is based on automated text-searches on the electronic archives of 10 newspapers: the Chicago Tribune, the Daily Telegraph, the Financial Times, The Globe and Mail, The Guardian, the Los Angeles Times, The New York Times, USA Today, The Wall Street Journal, and The Washington Post.
- The choice of six newspapers from the U.S., three from the United Kingdom, and one from Canada reflects author intention to capture events that have global dimension and repercussions

Geopolitical Risk Index (GPR)

Index Construction:

- GPR calculated as the share of articles discussing adverse geopolitical events:
- The index counts, each month, the number of articles discussing rising geopolitical risks, divided by the total number of published articles.
- By the same token, the historical GPR index, dating back to 1900, is based on searches of the historical archives of the Chicago Tribune, The New York Times, and The Washington Post.
- Separate sub-indexes for geopolitical threats and acts.
- To construct the outcome of interest, they use a dictionary-based method, specifying a dictionary of words whose occurrence in newspaper articles is associated with coverage of geopolitical events and threats.

Geopolitical Risk Index (GPR)

Index Construction:
The specification of information that guides the construction of the dictionary is the following:

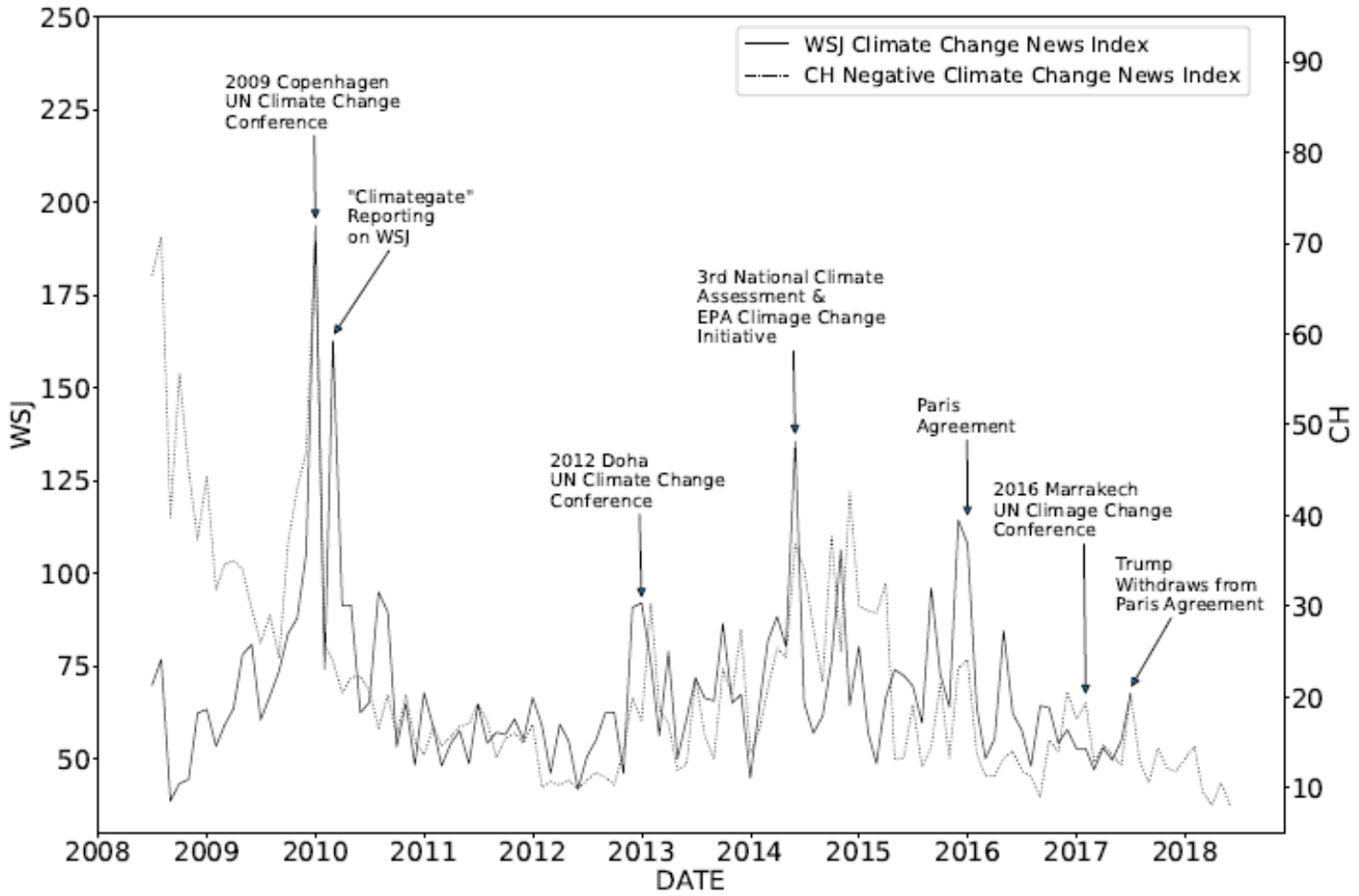
- First, they build directly on the definition of geopolitical risk adopted in this paper, selecting words that closely align with the definition.
- Second, they use information from two geopolitical textbooks and from the Corpus of Historical American English to isolate themes that are more likely to be associated with geopolitical events (such as 'war [on] terror' or 'nuclear weapon') or words that are more likely to be used in conjunction with war-related words (such as "declare").
- Third, they organize the search around high-frequency words and their synonyms that are more likely to appear in newspapers on days of high geopolitical tensions

Geopolitical Risk Index (GPR)

Index Construction:

- The goal is to provide an index that can highlight distinct aspects of geopolitical risk, and that can be sliced conceptually and geographically.
- Doing so exclusively with one-word searches would likely lead to misclassification and measurement error. These considerations lead the authors to a specific search query, which specifies two words or phrases whose joint occurrence likely indicates adverse geopolitical events.
- The query is described in the following slide, and is organized in eight categories . Each category is captured by a search query comprising two sets of words, the first set containing topic words (e.g. ‘war,‘ ‘nuclear,‘ or ”terrorism”), the second set containing ‘threat‘ words for categories 1 through 5 and ‘act‘ words for categories 6 through 8.

Geopolitical Risk Index (GPR)



This figure shows the CH Negative Climate Change News Index from 2008 to 2017, overlaid against the WSJ Climate Change News Index, and annotated with climate-relevant news announcements.

Geopolitical Risk Index (GPR)

A. Search Categories and Search Queries

	Category	Search Query	Peak (Month)	Contribution to Index%		
				Full sample	1900-1959	1960-2019
Threats	1. War Threats	War_words N/2 Threat_words	Germany Invades Czech. (September 1938)	13.5	17.9	9.2
	2. Peace Threats	Peace_words N/2 Peace_disruption_words	Iran Crisis of 1946 (April 1946)	3.5	4.3	2.7
	3. Military Buildup	Military_words AND buildup_words	Cuban Missile Crisis (October 1962)	23.5	21.3	25.8
	4. Nuclear Threats	Nuclear_bigrams AND Threat_words	Nuclear Ban Negotiations (August 1963)	10.1	4.2	16
	5. Terrorist Threats	Terrorism_words N/2 Threat_words	9/11 (October 2001)	2.7	0.3	5
Acts	6. Beginning of War	War_words N/2 War_begin_words	WWII Begins (September 1939)	18.8	26.8	10.7
	7. Escalation of War	Actors_words N/2 Actors_fight_words	D-Day (June 1944)	19.6	23.9	15.3
	8. Terrorist Acts	Terrorism_words N/2 Terrorism_act_words	9/11 (September 2001)	8.3	1.3	15.2

Geopolitical Risk Index (GPR)

B. Search Words

Topic Sets	Phrases
War_words	war OR conflict OR hostilities OR revolution* OR insurrection OR uprising OR revolt OR coup OR geopolitical
Peace_words	peace OR truce OR armistice OR treaty OR parley
Military_words	military OR troops OR missile* OR "arms" OR weapon* OR bomb* OR warhead*
Nuclear_bigrams	"nuclear war*" OR "atomic war*" OR "nuclear missile*" OR "nuclear bomb*" OR "atomic bomb*" OR "h-bomb*" OR "hydrogen bomb*" OR "nuclear test" OR "nuclear weapon*"
Terrorism_words	terror* OR guerrilla* OR hostage*
Actor_words	allie* OR enem* OR insurgen* OR foe* OR army OR navy OR aerial OR troops OR rebels

Threat/Act Sets	Phrases
Threat_words	threat* OR warn* OR fear* OR risk* OR concern* OR danger* OR doubt* OR crisis OR troubl* OR disput* OR tension* OR imminen* OR inevitable OR footing OR menace* OR brink OR scare OR peril*
Peace_disruption_words	threat* OR menace* OR reject* OR peril* OR boycott* OR disrupt*
Buildup_words	buildup* OR build-up* OR sanction* OR blockad* OR embargo OR quarantine OR ultimatum OR mobiliz*
War_begin_words	begin* OR start* OR declar* OR begun OR began OR outbreak OR "broke out" OR breakout OR proclamation OR launch*
Actor_fight_words	advance* OR attack* OR strike* OR drive* OR shell* OR offensive OR invasion OR invad* OR clash* OR raid* OR launch*
Terrorism_act_words	attack OR act OR bomb* OR kill* OR strike* OR hijack*

C. Excluded Words

Exclusion words	movie* OR film* OR museum* OR anniversar* OR obituar* OR memorial* OR arts OR book OR books OR memoir* OR "price war" OR game OR story OR history OR veteran* OR tribute* OR sport OR music OR racing OR cancer OR "real estate" OR mafia OR trial OR tax

Historical Insights of GPR

Key Historical Events:

- GPR spikes during major conflicts: World Wars, Korean War, Cuban Missile Crisis, 9/11.
- Recent spikes: Gulf War, 2015 Paris terrorist attacks, North Korea crisis.

Validation of the GPR Index

Validation Methods:

- Cross-referenced with military spending data and war deaths.
- Correlated with major newspaper front-page coverage to verify accuracy.
- High correlation with actual geopolitical events.
- Comparison with uncertainty (the VIX-a measure of stock market volatility-and the news-based EPU index of Baker, Bloom, and Davis (2016)).
- Finaly they evaluate the GPR index against alternatives based on different search queries and they perform an extensive human audit of newspaper articles likely discussing geopolitical risks.

Methodology in Detail

VAR Models:

- Assess macroeconomic impact of GPR.
- Includes real business investment, labor conditions, stock prices, and oil prices.
- main specification (using two lags and quarterly data from 1986:Q1 through 2019:Q4) consists of eight variables: (1) the log of the GPR index; (2) the VIX; (3) the log of real business fixed investment per capita; (4) the log of private hours per capita; (5) the log of the Standard and Poor's 500 index; (6) the log of the West Texas Intermediate price of oil; (7) the yield on two-year U.S. Treasuries; (8) the Chicago Fed's National Financial Conditions Index (NFCI).
- Separate models for geopolitical threats (GPT) vs. acts (GPA).

Impact of GPR on Macroeconomic Variables

Findings from VAR Models:

- High GPR leads to declines in investment, employment, and stock prices.
- Both geopolitical threats and realized acts contribute to economic contraction.
- Significant impact in highly exposed industries.

Firm-Level Data and Investment

Industry and Firm-Level Data:

- Analyzed 49 industry groups sensitive to GPR (e.g., defense, petroleum, transportation).
- Used earnings call transcripts to track firm-specific geopolitical risk.

Policy Implications

Policy Recommendations:

- Policymakers should incorporate GPR data into economic forecasting.
- Central banks and investors need to manage risks linked to geopolitical events.

Conclusion

Summary:

- Geopolitical risk significantly impacts investment, employment, and GDP growth.
- GPR index can help predict economic downturns, particularly in exposed industries.
- Future research can extend findings to non-English sources and other contexts.

Cybersecurity Risk

C. Florackis, C. Louca, R. Michaely and M. Weber

Introduction to Cybersecurity Risk

Understanding Cybersecurity Risk

- Cybersecurity risk refers to potential financial losses, disruptions, or reputational damage that firms may face due to failures in their information technology systems, often caused by cyberattacks.
- Examples of cybersecurity risks include data breaches, service disruptions, and physical damage to IT infrastructure.
- Firms are increasingly concerned about cybersecurity due to the rising frequency and complexity of cyberattacks.

The Research Goal

Developing a Cybersecurity Risk Measure

- The authors aim to create a novel firm-level measure of cybersecurity risk for U.S.-listed firms, which is based on textual analysis of risk disclosures.
- The goal is to explore whether cybersecurity risk is priced in the stock market and how exposure to cybersecurity risk correlates with future returns.

Machine Learning for Textual Analysis

Machine Learning Approach for Textual Analysis

- The paper utilizes **textual analysis** as the primary method for identifying cybersecurity risk from firms' **10-K filings**, specifically from the **Item 1A. Risk Factors** section.
- **Natural Language Processing (NLP)** techniques are used to extract language related to cybersecurity risk from these filings, which forms the basis of the machine learning model.

Data Collection and Processing

Data Sources

- **10-K Filings** from SEC's Edgar database, focusing on risk factors related to cybersecurity. The primary data comes from the "Item 1A. Risk Factors" section of 10-K filings, which is mandated by the U.S. Securities and Exchange Commission (SEC).
- Firms are required to disclose the most significant risks they face in this section, which makes it a valuable source for understanding how firms perceive and disclose cybersecurity risks. The authors focus on firms' descriptions of cybersecurity risk and use this information to create their risk measure.
- **PRC Database** for identifying firms that experienced data breaches or cyberattacks.
- Financial and stock return data from CRSP, Compustat, and Bloomberg.

Data Collection and Processing

Text Extraction

- They develop a list of keywords and phrases directly related to cybersecurity risks, such as "unauthorized access," "cyber," "data breach," "hacker," and "attack."
- To refine the extraction process, additional contextual filters are applied to ensure relevance. For example, they require keywords like "cyber" to be present along with "attack" to avoid irrelevant matches such as "terrorist attacks".
- This results in extracting sentences that specifically discuss cybersecurity risk.
- Examples of keywords: "unauthorized access", "cyber", "data breach", "hacker".

"Training" Sample of Firms Attacked by Cyberattacks

To create a reliable measure of cybersecurity risk, the authors begin by building a *training sample* of firms that have experienced major cyberattacks.

- **Privacy Rights Clearinghouse (PRC)** provides the data on firms that suffered cyberattacks between 2007 and 2018. PRC records the details of cyber incidents such as the type of attack (e.g., hacking, data breaches) and its impact.
- The training sample **consists of firms that have faced "major" cyberattacks**, meaning these incidents attracted significant attention in global news outlets like Bloomberg, Reuters, or major newspapers. These firms are expected to have had **high exposure to cybersecurity risk ex-ante**.

Indirect Descriptions of Cybersecurity Risk

In addition to direct mentions of cybersecurity risk, the authors also capture indirect descriptions.

- For instance, firms might describe their business operations, security measures, and the potential consequences of a cyberattack without explicitly using terms like "cyber" or "hacker."
- The authors categorize these descriptions into three consequences: internal consequences, legal consequences, and economic consequences.
- A second list of indirect keywords is compiled to capture these discussions, including terms like "damage," "disruption," and "litigation."

Similarity-Based Textual Analysis

- The crux of the cybersecurity risk measure is based on **textual similarity**. The idea is that firms with similar levels of cybersecurity risk will use similar language in their risk disclosures.
- **For each firm, the authors calculate the similarity between its cybersecurity-risk disclosure and the disclosures of firms in the training sample (firms that experienced major cyberattacks)**.
- Cosine Similarity and Jaccard Similarity are the two methods used for this comparison.

Similarity-Based Textual Analysis

a) Cosine Similarity

- Cosine similarity measures the cosine of the angle between two word vectors.
- The firm‘s disclosure is transformed into a vector of word frequencies. The similarity between two documents (in this case, a firm‘s 10-K disclosure and the disclosure of an attacked firm) is the cosine of the angle between the two vectors.
- A higher cosine similarity score indicates greater overlap in the language used, suggesting similar levels of cybersecurity risk.

Similarity-Based Textual Analysis

b) Jaccard Similarity

- Jaccard similarity measures the size of the intersection of two sets of words (shared words) divided by the size of their union (total unique words).
- It essentially captures the degree of overlap in word choice between two firms cybersecurity-risk disclosures.

These similarity measures help to quantify how closely a firm's risk disclosure resembles that of the attacked firms.

Resume The Machine Learning Process

Step 1: Text Extraction:

- Cybersecurity-related sentences are extracted from risk disclosures using a keyword-based algorithm.

Step 2: Feature Engineering:

- The sentences are transformed into word vectors, excluding stopwords, pronouns, and irrelevant phrases.

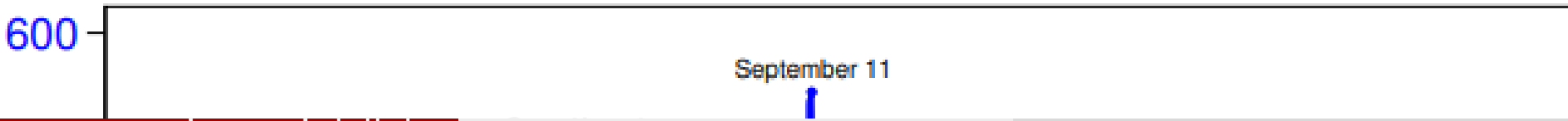
Step 3: Similarity Calculation:

- Each firm's textual data is compared with those in the training set using cosine and Jaccard similarity.

Step 4: Scoring:

- The similarity score for each firm is averaged to calculate its overall cybersecurity risk score.

Figure 1: Recent Geopolitical Risk Index from 1985



Validation of the Cybersecurity Risk Measure

The authors validate the cybersecurity risk measure through several tests:

- Language Complexity: Firms with higher scores tend to use more sophisticated and detailed language to describe cybersecurity risks.
- Legal Consequences: Higher-risk firms are more likely to discuss the legal consequences of cyberattacks and use more precise and negative language in their disclosures, potentially as a way to reduce litigation risk.
- Cyber Insurance: Firms with higher risk scores are more likely to mention cyber insurance policies in their filings, indicating they are actively managing their cybersecurity risk.

Application in Asset Pricing

- Finally, the authors apply the cybersecurity risk measure to study whether cybersecurity risk is priced in the cross-section of stock returns.
- They form portfolios of firms based on their cybersecurity risk scores and find that firms with higher cybersecurity risk tend to earn higher future returns, suggesting that investors demand a risk premium for holding stocks exposed to higher cybersecurity risk.
- This finding is further confirmed using Fama-MacBeth regressions, showing that cybersecurity risk is a significant predictor of stock returns.

Application in Asset Pricing

In particular ...

- Stock Returns: Firms with high cybersecurity risk scores tend to earn higher future returns, as investors demand a premium for holding these riskier stocks.
- Portfolio Performance: An equal-weighted portfolio of high-cybersecurity-risk firms outperformed low-risk firms by up to 8.3
- Cross-sectional Regression: Fama-MacBeth regressions show that cybersecurity risk predicts stock returns up to 12 months in advance.

Conclusion

The Role of Machine Learning in Understanding Cybersecurity Risk

- This study highlights the importance of machine learning techniques such as NLP and textual similarity measures in quantifying firm-level cybersecurity risk.
- The model demonstrates strong predictive power, correlating with actual cyberattacks and stock market performance.
- The approach provides valuable insights for investors, regulators, and firms about managing cybersecurity risk and its implications for financial markets.

The Anatomy of Cyber Risk

R. Jamilov, H. Rey and A. Tahoun

Objectives of the Study

- Introduce a novel measure of **firm-level** cyber-risk exposure using computational linguistics applied to transcripts from quarterly **earnings conference calls** from 2002-2021.
- These transcripts provide a direct insight into the discussions held between corporate executives and analysts or investors, often touching on a range of topics including financial performance, market challenges, and operational risks like cyber threats.
- Firm-level exposure correlates with actual cyberattacks impacting stock returns and profits. Market adjustments visible through equity options.

Source of Data

Scope and Scale

- **Firms**: The dataset encompasses approximately 13,000 firms. This broad coverage across different industries and sectors provides a comprehensive view of how businesses of various sizes and operational scopes discuss and perceive cyber risk.
- **Countries**: These firms are located across 85 countries, offering a global perspective on cyber risk that includes insights from both developed and emerging markets. The international diversity in the dataset is crucial for understanding regional differences in cybersecurity awareness and incident reporting.
- **Time Span**: The data covers the period from 2002 to 2021. This long timeframe allows the authors to analyze trends over nearly two decades, capturing the evolution of cyber risk discussions as digital technologies and cyber threats have developed.

Source of Data

Content of the Data

- **Earnings Calls**: The discussions during these calls generally revolve around the company's financial results, but they also delve into strategic issues like cybersecurity. This makes them rich sources of spontaneous, unscripted dialogue about the challenges and risks firms face, including those related to cyber threats.
- **Cybersecurity Discussions**: Specific mentions of terms and phrases related to cybersecurity within these calls were the focal points for analysis. These include direct mentions of cyber attacks, security breaches, IT security measures, and any other related discussions that could indicate a firm's exposure to cyber risk.

Data Extraction and Preparation

Transcript Acquisition:

- Transcripts were sourced from publicly available databases that collect and store such financial disclosures.
- In particular, quarterly earnings conference calls of firms are publicly listed in the United States from Thomson Reuters' StreetEvents

Option Market Data

- The main source of option data is the OptionMetrics' Ivy DB Volatility Surface File.

Text Processing:

- Before analysis, the text data would require significant preprocessing.
- This includes cleaning the data, tokenization and possibly tagging or categorizing terms based on their relevance to cybersecurity.

Analytical Use

Development of Cybersecurity Lexicon:

- The authors used these transcripts to develop a lexicon or list of terms that are relevant to cybersecurity, as discussed previously.
- This lexicon was then used to scan and score the earnings call transcripts for cyber risk discussions.

Validation Against Actual Incidents:

- To validate their cyber risk measure, the authors compared their findings from the textual analysis with actual reported cyber incidents and their financial impacts on the firms.
- This step is crucial to establish the reliability and relevance of the text-based measures derived from earnings calls.

Methodology

Development of a Cybersecurity Lexicon

A crucial step in their methodology was the creation of a comprehensive lexicon of cybersecurity-related terms. This lexicon includes terms that are commonly associated with cyber incidents, threats, and general cybersecurity practices. The development of this lexicon involved:

- Literature Review: Examining academic and industry literature to identify commonly used cybersecurity terms.
- Expert Consultation: Engaging with cybersecurity experts to validate and expand the list of terms.
- Regulatory and Industry Sources: Incorporating terms from regulatory frameworks and industry-specific documents that pertain to cybersecurity.

Methodology

Development of a Cybersecurity Lexicon Sources:

- Financial Stability Board (FSB) "Cyber Lexicon"
- "NCSC Glossary" of common cybersecurity terms provided by the National Cyber Security Centre
- "Glossary of Common Cybersecurity Terms and Phrases" made available by the NICCS, an initiative managed by the Cybersecurity and Infrastructure Security Agency (CISA)
-

In total, the dictionary used in the paper consists of 275 terms

Methodology

Natural Language Processing (NLP) Application

With the lexicon established, the next step involved applying NLP techniques to the earnings call transcripts. This process typically involves several stages:

- **Text Preprocessing**: Cleaning the text data by removing stopwords, punctuation, and performing tokenization to break down the text into manageable pieces or tokens.
- **Term Frequency Analysis**: Counting the frequency of each cybersecurity-related term within the transcripts to gauge the focus on cybersecurity in each earnings call.
- **Contextual Analysis**: Not just identifying the presence of cybersecurity terms but also analyzing the context in which these terms were used to ensure they relate to actual cyber risk discussions rather than coincidental mentions.

Scoring System

Quantifying Cyber Risk Exposure

The frequency and context of cybersecurity-related terms were used to create a cyber risk exposure score for each firm for each quarter. This scoring system likely involves:

- **Scoring Algorithm**: Developing a scoring algorithm that assigns weights to different terms based on their significance and the context of their usage.
- **Normalization**: Normalizing these scores across all firms and time periods to facilitate comparative analysis.

Scoring System

Quantifying Cyber Risk Exposure

The actual "scoring system" is an index constructed from the aggregated data of term frequencies adjusted by the context, it generally involves:

- Summing the weighted frequencies of cybersecurity terms.
- Adjusting for the firm size or industry where necessary to normalize the measure across different companies.
- Potentially time-weighting the terms to reflect more recent discussions more heavily, aligning with the evolving nature of cyber threats.
-

Scoring System

They define three variants of the same measure.

- First, absolute frequency ($CyberRisk^{A}_{i,t}$) which is the number of times terms from the validated dictionary that appear in each earnings-call transcript.
- Second, relative frequency ($CyberRisk^{R}_{i,t}$) which is $CyberRisk^{A}_{i,t}$ scaled by the total number of words in each transcript $B_{i,t}$.
- Finally, a binary indicator ($CyberRisk^{I}_{i,t}$) that takes the value of 1 if any of the terms in the validated dictionary appears in the transcript, and 0 otherwise

$$Cybersecurity\ Risk_{i,t} = \sum_{n=1}^{N} \frac{CS_{i,n,t}}{N_{t-1}}$$

Validation

Validation and Correlation with Actual Incidents

- To validate their NLP-derived measures of cyber risk, the researchers correlated these measures with actual reported cyber incidents and their impact on financial outcomes such as stock returns, profitability, and market valuation adjustments.
- This step is critical to ensure that the textual analysis provides a reliable indicator of real-world cyber risk.

Statistical and Econometric Analysis

Stock Market Effect

- The first test of economic significance is whether proposed measures of cyber risk exposure have any meaningful effect on firms‘ stock market performance.
- Cyber risk exposure does not necessarily imply an actual incident; it is fundamentally a forward looking measure which implies a heightened likelihood of a future cybersecurity crisis or event. This uncertainty alone can affect asset prices today.
- To test this theory, they run quarterly firm-level regressions of standardized value-weighted stock returns (WRet), cumulative stock returns (CRet), and realized stock market volatility (RV) on CyberRiskIi,t and CyberRiskRi,t
- Both have negative and significant effects on stock returns and have a large and significant positive effect on realized volatility