Extracting geopolitical risk indices from news data with sentiment analysis

University of Turku
Department of Computing
Master of Science in Technology Thesis
Master's Degree Programme in ICT (Data Analytics)
June 2025
Aleksi Pikkarainen

UNIVERSITY OF TURKU
Department of Computing

Aleksi Pikkarainen: Extracting geopolitical risk indices from news data with sentiment analysis

Master of Science in Technology Thesis, 45 p.
Master's Degree Programme in ICT (Data Analytics)
June 2025

Geopolitical risk has been measured and quantified all the way since the 1950s. Over time, technologies to model it have evolved and various methods have been employed to construct geopolitical risk indices. These indices give insight to ongoing conflicts, trade wars and trade agreements between nations. Lately, utilizing natural language processing to create real-time indices from various internet data sources has been a rising trend. In today's constantly changing world and overabundance of information acquiring critical information through these indices is extremely beneficial for those in business and governmental decision making.

This thesis explored geopolitical risk indices through constructing a pipeline that utilizes various natural language processing methods. I researched what kind of elements this pipeline requires, what are the most efficient ways to extract information and how the end result can be shaped into a time series index. Mission Grey's News Dataset was used to both act as a source for the indices and to provide annotatable raw data for used information extraction methods.

The results showed that using a transformer-based text classifier could sufficiently categorize news data based on geopolitical context. Named entity recognition was used in combination to detect which countries these news articles discuss. A fine-tuned sentiment analysis model was an efficient way to extract polarity from chosen articles. This extracted polarity was transformed into indices using four different methods. From these methods, the total news ratio -method showed performance close to the state-of-the-art geopolitical index.

Keywords: geopolitical risk, sentiment analysis, text classification, named entity recognitiona

Contents

List of Figures

List of Tables

List of acronyms

GPR Geopolitical Risk

GPT Generative Pre-trained Transformer

IRF impulse response function

KNN K-Nearest Neighbours

LLM Large Language Model

NER Named Entity Recognition

NLP Natural Language Processing

OLS Ordinary Least Squares

POS Part-of-Speech

RNN Recurrent Neural Networ

SVM Support Vector Machine

VAR Vector Auto-Regression

WEIS World Events Interaction Survey

1 Introduction

Geopolitics has an important effect on the total economy of the world [1]. This total economy is continuously growing, having reached a total value of 122 trillion dollars back in 2022 [2]. This is influenced by conflicts and war, which have a major effect on exports from developing countries. With the start of the Russian-Ukraine war, grain and crop exports from both countries were estimated to cause significant economic impacts [3] and increase global food insecurity [4]. The Israeli-Palestine conflict was deemed to affect the global energy and investment markets [5]. Alongside these major geopolitical events, the COVID-19 pandemic left a severe impact on various industries, economies, and assets [6]. For stakeholders and governmental decision makers, being able to stay up-to-date on geopolitics is a vital asset. Furthermore, being able to predict future large-scale issues such as these would be extremely beneficial.

Detecting and predicting these global geopolitical issues is difficult and resource intensive. Geopolitical risk (GPR) has been tried to quantify and analyze throughout the 20th century [7]–[12]. To alleviate this and better follow the changing global landscape, GPR indices have been used to describe the overall geopolitical situation of the world and countries. In the 2000s, advances in natural language processing (NLP) and machine learning have made it possible to extract geopolitical information from various text data more efficiently. The internet is full of social media posts and news, which provides an excellent source to obtain information about the current state of the world in textual form real-time.

To filter textual data suitable for the construction of GPR indices, keyword-based methods have been popular among practitioners. They allow an easy way to quickly select

relevant data. However, this requires a lot of domain knowledge and manual work. The language centered around geopolitics changes over time, so keywords and phrases need to be updated regularly to ensure accurate filtering. Creating a method that would not require much domain knowledge and manual work and could then be adjusted for the new language would be a valuable asset in creating GPR indices. In addition, extracting the underlying information from selected news articles has mostly centered around increased GPR. Could topics that lower the overall GPR be also taken into account when creating such indices?

In this thesis, I propose a new pipeline approach to create real-time GPR indices based on news data. Using the transformer architecture, geopolitical data can be filtered more efficiently and modified when needed. This enables the pipeline to nearly independently continue producing new weekly GPR indices by country as new news data comes in. I explore my approach with the following two research questions:

RQ1: What kind of components are needed to transform news data into GPR indices?

RQ2: How can sentiment be transformed into indices?

For RQ1 I explore in-depth what kind of methodologies I need to use in order to efficiently transform constant news data into a weekly GPR index. The aim of this paper is to address this question by selecting applicable methods for each task and evaluating their performance. The selected methods need to be able to generalize for new, unseen articles and perform their tasks independently with minimal human intervention. For RQ2 I study and test different methods to transform sentiment resulting from my analysis into indices. These indices are then compared to each other. The most fitting index is selected to be compared to a prominent GPR index.

My thesis consists of 7 chapters. In chapter 2 I analyze the existing research and methodologies done in the field of geopolitics and GPR index extraction. In chapter 3 I explain the NLP methodologies I will be using as a part of my pipeline. In chapter 4 I present the Mission Grey News Dataset and annotated datasets I will be using for training components of my pipeline. In chapter 5 I will go through the actual pipeline

method creation. In chapter 6 I will present the results acquired from chapter, answer the research questions and reflect on our findings. Chapter 7 is reserved for discussion about possible directions future research could take.

2 Literature review

In this literature review, I go through some of the earlier research exploring how to process textual data into GPR indices. Afterwards I discuss recent examples which have leveraged the use of natural language processing for this task and from which I borrow methodologies to aid my research.

The first major milestone in attempting to derive indices from textual data could be attributed to Joshua S. Goldstein, who proposed a conflict-cooperation scale based on World Events Interaction Survey (WEIS) [7], [12]. In his paper he critized the past works of Bobrow [9], Dixon [10] and Ward [8] for inconsistencies on defining categories to label these world events on. Goldstein wanted to better highlight the weighting aspect of different WEIS categories, rather than counting the raw amount and ratio of different categories occurring within the events data. To create weights for each of the WEIS categories, he employed a diverse panel of eight faculty members of the University of Southern California to label 61 different WEIS event types. They were given cards for each different event type and asked to sort them into three different categories: cooperative, conflictual and neutral actions. Then, within each category the cards were then placed into 10 different ordered boxes, which each represented the magnitude of the event type as box 0 representing the most neutral, and 10 the most conflictual act among them. To maintain linearity, the categories were compared in a way so that for example, box number 4 contained events twice as conflictual as those in box number 2. Afterwards, the event types were weighted based on which box their card ended up in, with cooperative events being labeled with a "+" and conflictual events with a "-" (table 2.1). Goldstein evaluated his conflict-cooperation scale comparing monthly time series created by it to

the Vincent scale [13]. Overall, his approach of weighting geopolitical keywords became a solid foundation for future research to build on.

Table 2.1: Goldstein weights for WEIS Events [12]

Code	*Event Type*	*Weight (SD)*
223	Military attack; clash; assault	−10.0 (0.0)
211	Seize position or possessions	−9.2 (0.7)
222	Nonmilitary destruction/injury	−8.7 (0.5)
221	Noninjury destructive action	−8.3 (0.6)
182	Armed force mobilization; display; buildup	−7.6 (1.2)
195	Break diplomatic relations	−7.0 (1.3)
173	Threat with force specified	−7.0 (1.1)
174	Ultimatum; threat w/ neg. sanction	−6.9 (1.4)
172	Threat with specific nonmil. sanction	−5.8 (1.9)
193	Reduce aid; punish/deprive	−5.6 (1.4)
181	Nonmil. demonstration, walk out	−5.2 (2.1)
201	Order personnel out of country	−5.0 (1.7)
202	Expel organization or group	−4.9 (1.4)
150	Issue demand; insist compliance	−4.9 (1.7)
171	Threat w/o specific sanction	−4.4 (1.5)
212	Detain or arrest person(s)	−4.4 (2.3)
192	Recall officials; reduce contact	−4.1 (1.2)
112	Refuse; oppose action	−4.0 (1.5)
111	Reject protest, demand, threat	−4.0 (1.5)
194	Halt negotiation	−3.8 (0.9)
122	Denounce; abuse	−3.4 (1.1)
160	Give warning	−3.0 (1.3)
132	File formal protest	−2.4 (0.9)

Note: Weight is mean of weights assigned by eight panelists; *SD* is standard deviation.

Table 2.1 Continued

Code	*Event Type*	*Weight (SD)*
121	Criticize; disapprove	−2.2 (1.3)
191	Cancel planned event	−2.2 (1.5)
131	Make informal complaint	−1.9 (0.6)
063	Grant asylum	−1.1 (2.5)
142	Deny policy/role/position	−1.1 (1.0)
141	Deny accusation	−0.9 (1.3)
023	Comment on situation	−0.2 (0.5)
102	Urge action or policy	−0.1 (1.5)
021	Decline to comment	−0.1 (0.6)
094	Request action	−0.1 (1.0)
025	State future policy	0.0 (0.0)
091	Ask for information	0.1 (0.4)
011	Surrender; yield to order	0.6 (7.2)
012	Yield; retreat; evacuate	0.6 (6.6)
031	Meet; send note	1.0 (0.9)
095	Plead; appeal; beg	1.2 (1.8)
101	Offer proposal	1.5 (1.9)
061	Express regret	1.8 (1.5)
032	Visit; go to	1.9 (2.4)
066	Return persons or property	1.9 (2.7)
013	Admit wrongdoing	2.0 (2.2)
062	State invitation	2.5 (2.7)
054	Assure; reassure	2.8 (2.2)
033	Host visit	2.8 (3.0)
065	Suspend sanctions; truce	2.9 (3.6)
082	Agree to future negotiation	3.0 (2.5)

Note: Weight is mean of weights assigned by eight panelists; *SD* is standard deviation.

Table 2.1 Continued

Code	*Event Type*	*Weight (SD)*
092	Ask for policy help	3.4 (1.1)
093	Ask for material help	3.4 (2.4)
041	Praise; condolences	3.4 (2.1)
042	Endorse verbally	3.6 (1.8)
053	Promise future support	4.5 (1.6)
051	Promise own support	4.5 (1.7)
052	Promise material support	5.2 (1.5)
064	Diplomatic recognition	5.4 (1.4)
073	Give other aid	6.5 (1.9)
081	Make agreement	6.5 (1.4)
071	Give economic aid	7.4 (1.0)
072	Extend military aid	8.3 (0.9)

Note: Weight is mean of weights assigned by eight panelists; *SD* is standard deviation.

In 2000, Feng built on Goldstein's research and proposed an approach to measure cross-country conflict based on time-series data [14]. He used three event datasets: WEIS, COPDAB [15] and KEDS [16]. Using ordinary least squares (OLS) regression and statistical stability testing, he was able to combine both WEIS and COPDAB in a comparable manner to produce a new dataset: COPWEIS. On this new dataset, he performed vector auto-regression (VAR) and impulse response function (IRF) analysis to explore dependencies between events from three major countries: the Soviet Union, Great Britain and the United States (fig. 2.1). From this analysis, he was able to use trade and conflict within these nations as dependent variables to explore these complex, non-linear and dynamic relationships based on these two variables.

VAR analysis to forecast possible conflicts was explored more in 2011 with Bayesian VAR models, when Brandt, Freeman and Schrodt proposed a framework to produce near

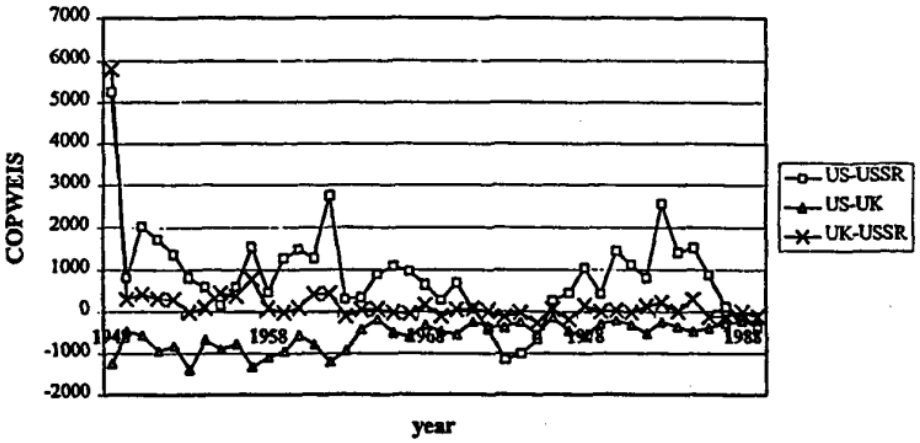

Figure 2.1: Feng's integrated Net Weighted Conflict Score for the United States, the United Kingdom amd the USSR, 1948-1989 [14]

real time time series predictions of political conflict between and within nations [17]. This framework utilized the automatic coding program, TABARI, in conjunction with the CAMEO event coding system to automatically acquire event data from international media sources such as Reuters and Agence France Press. This event data would then be passed to estimator models, which could continuously provide density evaluations of forecasts in real-time (fig. 2.2). They highlighted three major challenges which are still relevant today. Firstly, political actors, parties and terminology change over time. The vocabulary used to automatically select relevant event data has to be maintained over time in order for accurate forecasts to be calculated. Secondly, filtering news that could cause major misclassifications needs to be taken into account. In media, sports and entertainment news use terms similar to geopolitical news which can affect the forecasts. Having duplicate news is another contributor to these misclassifications, especially when using multiple different sources of news data. Thirdly, depending on the source, region or editorial policy the extensiveness of news coverage might vary. This can have an effect on the forecasts, especially when trying to predict areas with less coverage.

Some more recent examples include Caldara and Iacoviello [18], who proposed a GPR index constructed by selecting a dictionary of words related to geopolitics and GPR and analyzing 25 million news articles between 1900 and 2022. The method they utilized was comparing the ratio of news articles discussing GPR to the total number of published

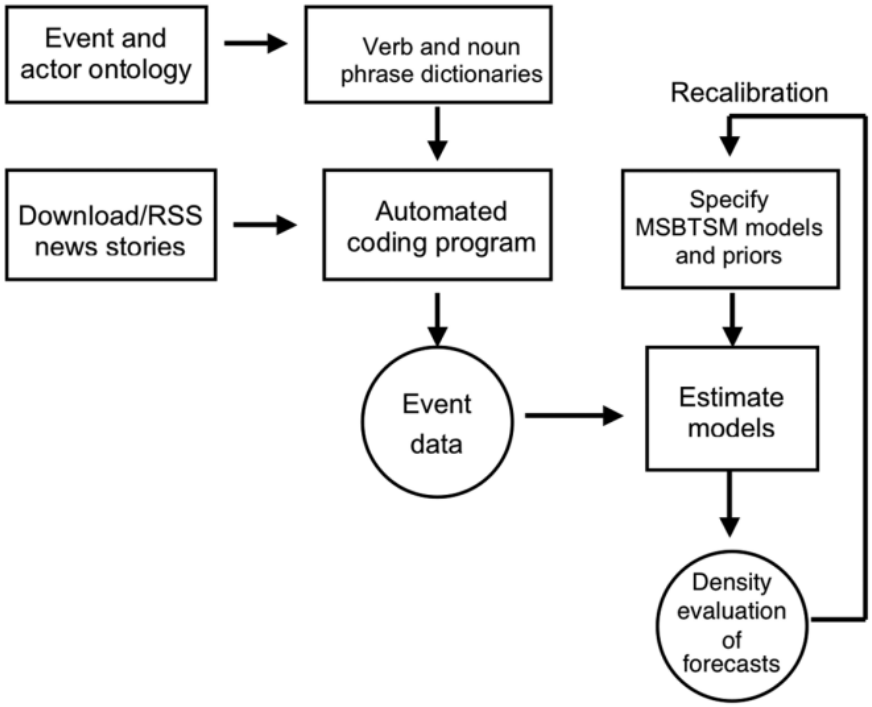

Figure 2.2: Data generation, model specification and forecasting framework [17]

news articles. They used a dictionary-based method by selecting specific words aligning with their definition of GPR and constructed bigrams to avoid possible miscalculations. These bigrams contained words belonging to eight distinct categories: war threats, peace threats, military buildup, nuclear threats, terrorist threats, trade war threats, beginning of war, escalation of war and terrorist acts. By utilizing these categories, which could be grouped into threats and acts, they were able to further divide this index into separate Geopolitical Threats index and Geopolitical Acts index by selecting which group of categories the resulting articles belonged to. This GPR index was validated by a few different methods. Firstly, a narrative GPR index was constructed by analyzing the New York Times headlines and assigning values to them depending on if they implied geopolitical risk on a scale of 0-5. These values are then used to construct a comparable index to the GPR index. The main GPR index is additionally compared to war deaths, news about military spending, proxies for uncertainty and granger causality tests for a more comprehensive validation. Through these detailed validation exercises they are able to prove that this GPR index accurately captures intensity and timing of adverse geopolitical events within countries and over time.

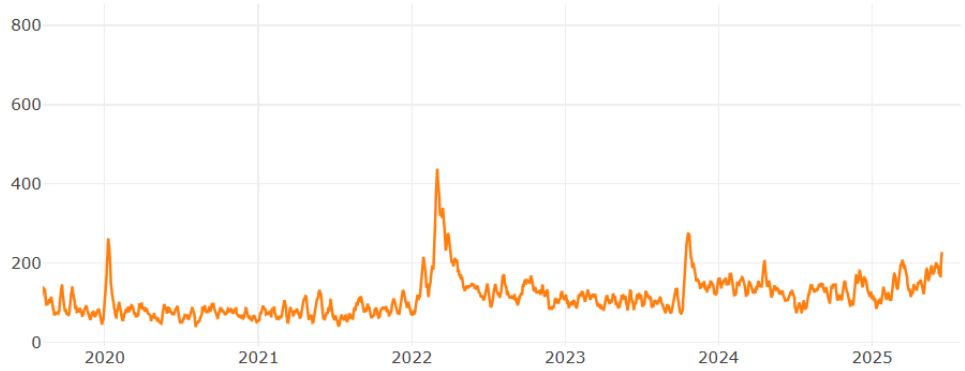

Figure 2.3: Caldara and Iacoviello's Weekly GPR index 2020-2025 [18]

Burns (2024) took this a step further by proposing multiple approaches to utilize textual machine learning methods to extract real-time GPR data from Twitter/X data. In his first case study, he took the Goldstein Index [12] as a base and split the event types used in it into singular words and bigrams. Using this vocabulary, he collected tweets discussing GPR themes. With this data, he utilized sentiment analysis to obtain daily changes in sentiment and constructed a time series based on it. This time series could then be analyzed to see if any major known geopolitical events, such as the Ukraine-Russian War, was visibly noticeable on it. In two time series: the daily sums of tweets containing terms from his Goldstein dictionary and the daily sum of sentiment, a massive spike (fig. 2.4) around the 24th of February 2022 when the Russian invasion on Ukraine started suggests that both keyword and sentiment analysis methods can be used to detect GPR from social media data (fig. 2.5). Additionally in this case study, Granger causality was used to explore whether or not this conflict had an effect on the financial markets 2.2. While the effects were not imminent, long-term visualizations showed that there was financial news sentiment that could be used to predict economic developments. In his second case study he explored the possibility of recording real-time changes as emerging geopolitical topics on Twitter / X by utilizing multilingual topic modeling. This would enable governmental decision makers and other important actors to respond rapidly to future conflicts. Importance of multilinguality is highlighted due to all events not having global implications and therefore would be visible only in regional news data while still being important to be detected. Burns used papers of Caldara and Iacoviello's on

geopolitical risk [18] and Klement's on geopolitics [19] to choose keywords and phrases to aid his topic modeling process. To perform the topic modeling, LDA algorithm [20] was used and the results were evaluated by comparing the generated topic trend lines to Google trends data (fig. 2.6). The data used in topic modeling was recorded between February 4th, 2023 and March 23rd, 2023. During this time, a significant geopolitical risk event happened when North Korea launched missiles in response to US and South Korea military exercises, conducted jointly nearby. This prompted spikes in the emerging topics four times out of the ten missiles launched and two of the spikes on X/Twitter happened two days before the Google Trends spike (fig. 2.7). With his results and prior, similar research done by Rill et. al [21] Burns could prove that emerging geopolitical topics could be followed near real-time using multilingual social media data and topic modeling. In the third case study Burns proposes a sentiment analysis approach to monitor GPR multilingually in real-time by combining aspects from his two earlier case studies. Using the Twitter/X API streaming, he is able to obtain social media post data continuously and the geopolitical bigrams he created in Case study 1 acted as a filter to detect which posts contained geopolitically relevant topics. For multilingual sentiment analysis he utilizes a multitude of methods such as VADER sentiment lexicon [22] for English, CamelBERT [23] for Arabic and custom recurrent neural networks (RNN) for French, Japanese, Korean, Portuguese and Spanish. Through these methods Burns managed to create a time-series sum of sentiment graph for all of these language's tweets combined which he could then compare to Goldstein index and his Case Study 1 index (fig. 2.8). Changes in sentiment trends are also checked for correlation with financial market terms, with results of no direct correlation.

In my thesis I propose a similar approach to Burns with using sentiment analysis to extract GPR from textual data. There are two differences with our research: I utilize news data instead of Twitter/X data to perform my analysis and I cannot perform multilingual analysis due to the dataset provided to me being in English. Additionally compared to many previous methods which utilized keywords to filter data, I face the issue of changing geopolitical vocabulary by utilizing machine learning methods instead of manual keywords

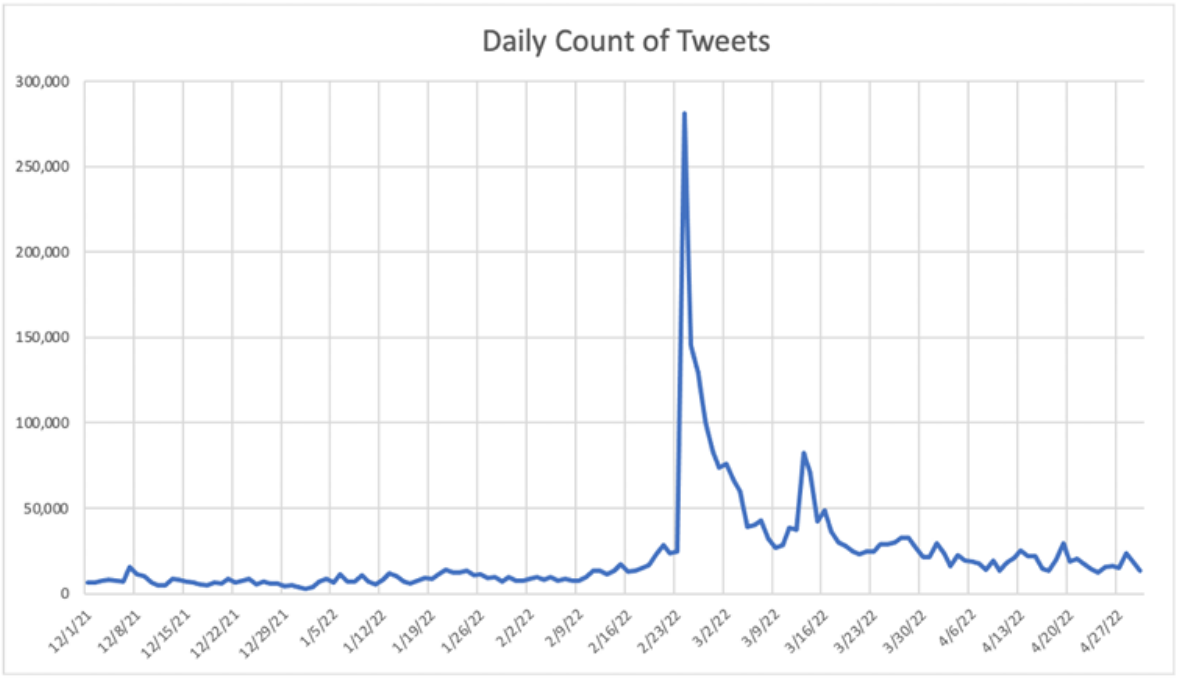

Figure 2.4: Burns' daily count of tweets in both "Goldstein positive" and "Goldstein negative" bigram categories[24]

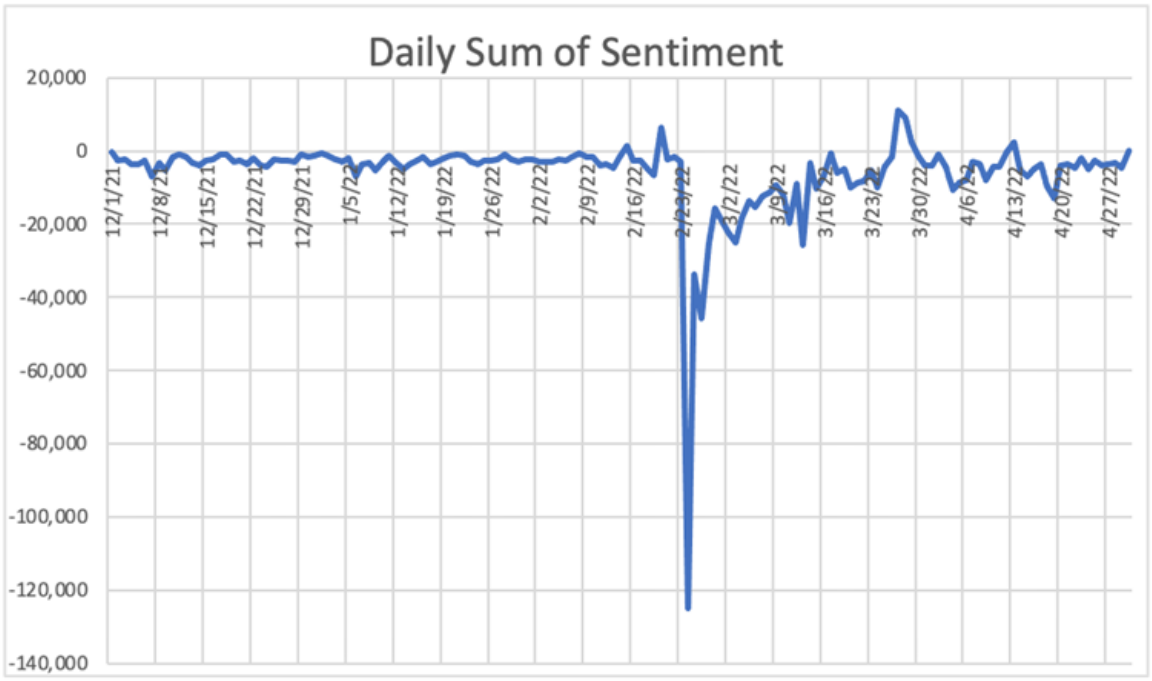

Figure 2.5: Burns' daily sum of sentmiment of all tweets in both "Goldstein positive" and "Goldstein negative" bigram categories[24]

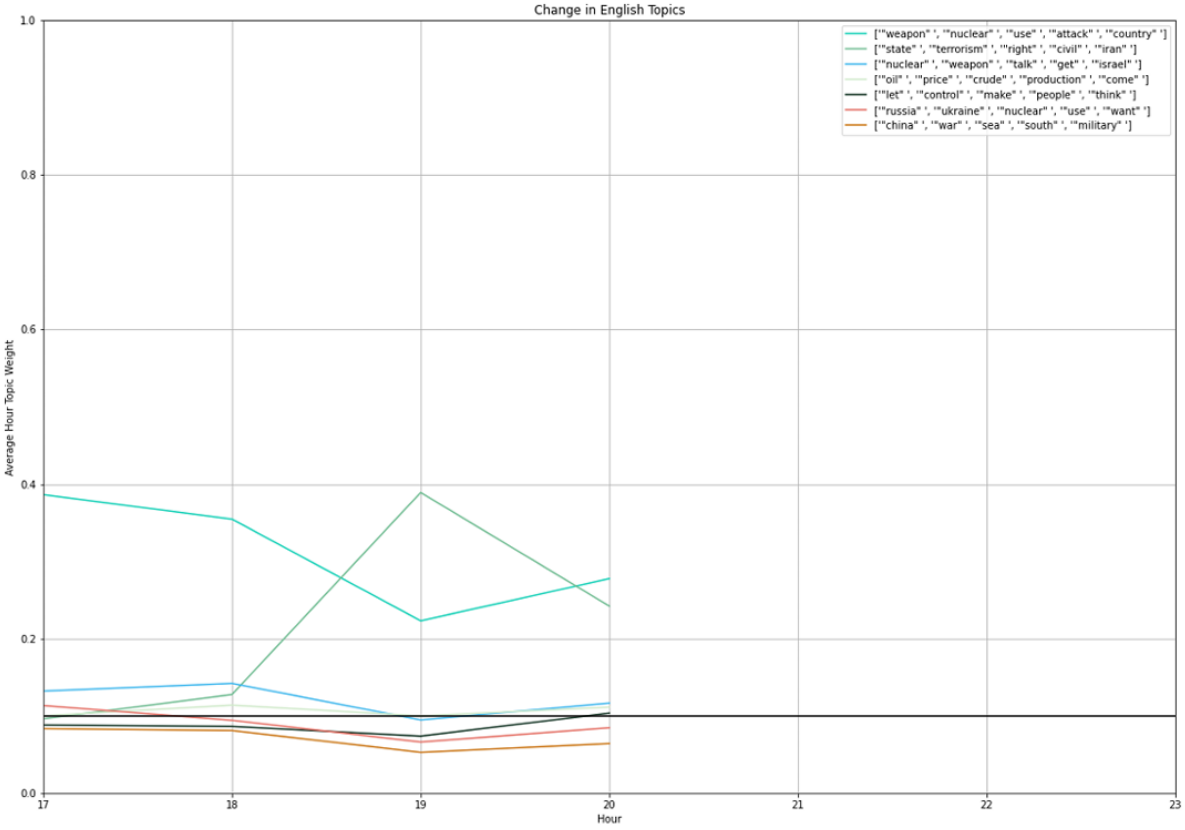

Figure 2.6: Burns' english topics that appeared and were tracked over time for June 1st, 2023[24]

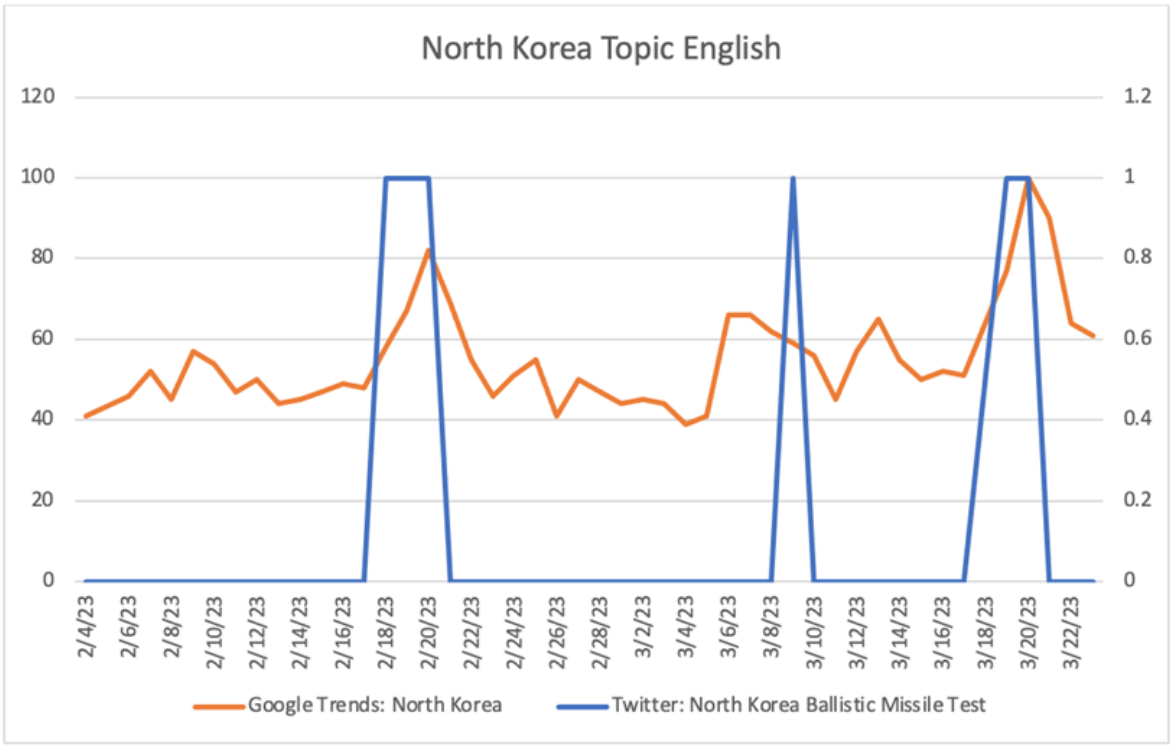

Figure 2.7: Burns' comparison between the Twitter/X emerging geopolitical topics data for English and Google Trends search data for North Korea. In the legend the Google Trends label is the criteria used to gather the Google Trends data. The Twitter / X legend label is the topic label used for the emerging topic on Twitter/X[24]

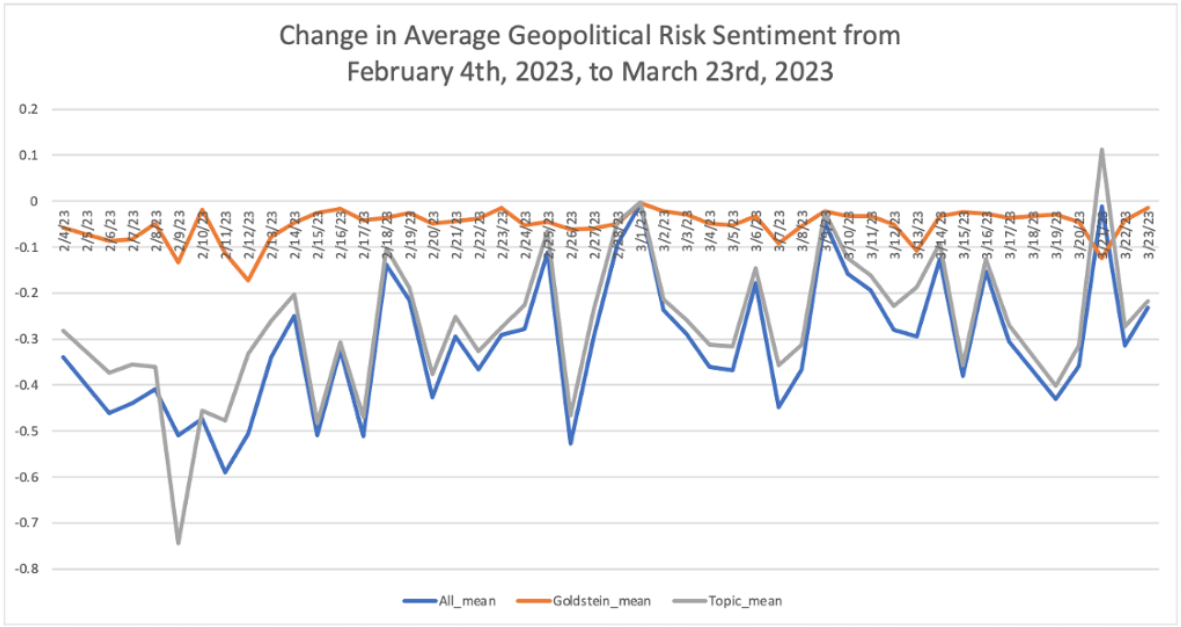

Figure 2.8: The change in the average sentiment for the geopolitical risk groupings on Twitter / X, with the blue line is "All", the orange line is "Goldstein", and the silver line is "Topics"[24]

Granger Causality	SentSum	PValue	Reverse Okay	Reverse PValue
Gold Price	Yes	0.003813	Yes	0.489212
Oil Price	Yes	3.09826E-07	Yes	0.936389
Gold Futures	Yes	0.000121234	Yes	0.969781
Oil Futures	Yes	8.57092E-08	Yes	0.927111
Wheat Futures	Yes	1.3972E-14	Yes	0.608716
Nikkei 225	Yes	0.032658	Yes	0.680651
German 10y Bond	Yes	0.000445	Yes	0.826079
FTSE 100	Yes	0.001059	Yes	0.806077
10 Year US Treasury	Yes	0.000575	Yes	0.23006
Defense-ETF	Yes	0.005718	Yes	0.803637
Metals-ETF	Yes	0.039279	Yes	0.977376
10 Year US Futures	Yes	0.000117	Yes	0.212961
Bitcoin-Futures	Yes	0.010613	Yes	0.727189
EUR	Yes	0.000284	Yes	0.984743
GBP	Yes	0.000997	Yes	0.943754
MXN	Yes	0.000004	Yes	0.857881
RUB	Yes	0.000002	Yes	0.98621
Bitcoin	Yes	0.004504	Yes	0.214014

Table 2.2: Burns' Granger Causality Results with Sentiment Summary [24]

and bigrams. I compare the effectiveness of an encoder-based text classifier transformer to a keyword corpus generated by a decoder-based large language model (LLM). I extract country information from my news data similarly to Caldara and Iacoviello [18], who utilized country name occurrences within the data to determine which country the data articles were related to. I streamline this process by proposing a named entity recognition (NER) approach to extract country information from news articles.

3 Technical background

Before I move on to the research, I explain in detail some of the key concepts used in both past and my current research. These concepts include sentiment analysis, sequence labeling, named entity recognition, text classification and commonly used machine learning evaluation metrics.

3.1 Text classification

Text classification is said to be: "the most fundamental and essential task in natural language processing" [25]. In its simplest form, it could be described as: "Assigning a string of text a label from a set of predetermined labels." The first statistical methods to perform it are all the way from the 1900s, such as K-Nearest Neighbours (KNN) [26] and Support Vector Machines (SVM) [27]. As computing power has increased during the 2010s and more research has been done, deep learning with neural networks and transformers have far outperformed the old, manual methods, especially in tasks such as sentiment analysis [28], question answering [29] and topic labeling [30]. In my thesis, I compare two approaches to classifying texts based on if they contain geopolitical information or not: a BERT[31]-based transformer model and a keyword-based method by utilizing a generative pre-trained transformer (GPT)[32] model.

3.1.1 Sentiment analysis

Sentiment analysis is a text classification technique used in NLP to extract sentiment from text 3.1. Sentiment can refer to a multitude of concepts falling under this term,

Table 3.1: Example sentiment analysis

text	label
This is the best movie I have seen all year! I really recommend this to all horror movie lovers!	1: Positive
I have never been so bored while watching a movie. Avoid this one at all costs!	0: Negative

but the most common is polarity, whether the overall tonality of the text is positive, neutral or negative. This can be achieved by marking individual words by either positive or negative and comparing them to the context they appear in the sentence [33]. These can be used to extract the overall sentiment from the text. Sentiment analysis is proven to be an efficient method to extract tonality from news article texts. Taj et. al. proposed the approach to use it to observe the polarities of five different classes of news: business, entertainment, politics, sports and tech [34].

In this thesis I explore the possibilities of using this extracted tonality to detect whether or not geopolitical news articles increase or decrease geopolitical risk. I am using a distilBERT-based robust-sentiment-analysis model [35] which I fine-tune to geopolitical news data to perform my analysis. I chose this model due to its lightweightness, good general performance and it having five classes instead of the commonly occurring three. These are the two additional tonalities, "very negative" and "very positive" that allow me to better capture broader sentiment present in geopolitical news articles.

3.2 Sequence labeling

Sequence labeling, often referred to as "sequence tagging" and "token classification", is a similar task to text classification, but instead of aiming to classify whole texts it is used to classify smaller spans of texts. Tasks involving sequence labeling are such as Part-of-Speech (POS) tagging where each word is assigned a part-of-speech tag such as noun or verb, and NER where multiple token spans forming mentions of names, locations and organizations are tagged with the corresponding labels. Neural networks are often used for these tasks due to their great performance and low need for domain expertise [36].

3.2.1 Named entity recognition

NER is a Sequence Labeling task that involves marking named entities in text. The marked entities usually belong to predetermined categories of interest, such as persons, organizations and locations. With entities consisting of multiple tokens BIO-tagging is used to indicate where the entity starts and ends. Entity starting points are marked with "B", the following tokens belonging to it as "I" and non-entities as "O" 3.2. I utilize NER to locate which countries the geopolitical news articles mention. Past research has utilized a similar method with keywords [18]. I use a BERT-based fine-tuned named entity recognition model dslim/bert-base-NER [37] with state-of-the-art performance.

Table 3.2: Example NER with BIO-tagging

Token	**Tag**
Donald	B-PER
Trump	I-PER
works	O
at	O
the	O
White	B-LOC
House	I-LOC
in	O
the	O
USA.	B-LOC

3.3 Machine learning evaluation metrics

Evaluation metrics are used in analyzing how well model predictions match the gold-standard predictions. They are based on simple formulas and utilize the four prediction quadrants found in a confusion matrix 3.3: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True positive is the amount of positively predicted, actual positive data points. True negative is the amount negatively predicted, actual negative data points. False positive is the amount of positively predicted, but actual negative data points. False negative is the amount of negatively predicted, but actual positive data points. All machine learning evaluation metric scores range from 0 to 1 which signifies the fraction each metric labeled correctly. 0 can be interpreted as the

Table 3.3: Confusion matrix

		Predictions	
		1	0
Actual	1	TP	FN
	0	FP	TN

model predicting wrong every time and 1 as the model predicting correctly every time. 0,5 is often referred to model "guessing" and depending on task, 0,8 to 0,95 is often a desirable score to achieve. These can vary depending on the metric, as some metrics can receive a value of 1 if the model labels every data point as "correct".

3.3.1 Accuracy

Accuracy is the rate of correct predictions, both positive and negative. It can be calculated by dividing the sum of all predictions from the sum of all correct predictions (true positives and true negatives). This produces the accuracy formula 3.1.

Accuracy metric:

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Out of all four metrics, accuracy provides the least useful information. This is due to it taking account of the amount of true negative predictions. If the data contains only a few positive values, the model marking every option as negative receives a high accuracy. With predictions, we are rarely interested in the true negatives. The most important aspect of a model is to understand how many times it is actually correct compared to the gold-standard predictions, so the next two metrics that compare the ratios of true positives prove to be useful in that regard.

3.3.2 Precision

Precision takes into account only the number of correct positive predictions, comparing them to the incorrect positive predictions. It can be calculated by dividing the sum of true positives and false positives from the amount of true positives. This produces the precision formula.

Precision metric:

$$\frac{TP}{TP + FP} \tag{3.2}$$

Precision is especially useful with an imbalanced class distribution and when the task wants to minimize false positives.

3.3.3 Recall

Recall takes into account only the number of correct positive predictions, comparing them to the incorrect false predictions. It can be calculated by dividing the sum of true positives and false negatives from the true positives. This produces the recall formula 3.3.

Recall metric:

$$\frac{TP}{TP + FN} \tag{3.3}$$

When we do not care about false positives and instead want to minimize false negatives i.e. predicting as many correct gold-standard predictions as correct as we can, we should prioritize recall over precision. With recall and precision, most of the time increasing the other decreases the other. Depending on the task, one should be prioritized over the other if possible.

3.3.4 F1-score

F1-score is a balanced mean between recall and precision. It can be calculated by dividing the sum of precision and recall from the multiplied precision and recall. Then the resulting ratio is multiplied by two. This produces the F1-score formula 3.4.

F1-score metric:

$$2 * \frac{precision * recall}{precision + recall} \tag{3.4}$$

With the F1-score balancing both precision and recall, it is a good, general metric to use when you're not focused on either precision or recall. It can be used as a great starting point to then start analyzing precision and recall from, since they present more

intricate information about the model predictions altogether. Having both high precision and recall can be difficult to achieve since increasing other often results in the other decreasing. A high F1-scoring model can be assumed to have an excellent performance.

4 Data

Mission Grey, a company offering AI solutions for global business decision making, provided me with a dataset to use for this research. It contains English news data from Google News between the years of 1984-2024 and I am using it to train my models, evaluate my models and extract the indices from it. One article datapoint contains an article id, country it mainly refers to, publisher link, short description, publication date, title, link to article and article text. The data is generally filtered to be geopolitically related through Google News tags, but in order to ensure only geopolitically relevant articles are taken into account, I manually create annotated data that can be used to fine-tune classifiers and validate classification approaches. The articles are annotated based on if they are related to geopolitics or not. I will next go into more detail about the dataset as a whole and then discuss the annotated data and annotation process.

4.1 Mission Grey News Dataset

The Mission Grey News Dataset contains 1104194 different news articles, with publication dates ranging from 16th of August 1984 to 21st of October 2024. The number of articles per year is greatly skewed towards 2024. These news articles, acquired from Google News, represent a wide variety of different countries and newspapers. The data was acquired through the live feed in October of 2024, with an attempt to select only the most recent news, which explains the skewness found within as seen from the table 4.1

Exact information about the method to filter them is not available, but the "country" column suggests that they were chosen by a country-based filter. This is further reinforced by each article containing at least one country in the article text. This however causes

Table 4.1: Article distribution by publication year

Year	Articles	Year	Articles
1984	2	2011	44
1986	1	2012	67
1999	1	2013	80
2000	1	2014	116
2001	2	2015	173
2002	12	2016	217
2003	3	2017	305
2004	6	2018	435
2005	12	2019	477
2006	8	2020	564
2007	13	2021	812
2008	7	2022	1116
2009	23	2023	18752
2010	30	2024	1080915

Table 4.2: Article distribution by month 2023-2024

	Year	
Month	2023	2024
1	92	7008
2	91	7445
3	154	17959
4	171	146231
5	462	183793
6	859	215596
7	1161	119832
8	1607	208654
9	2691	100967
10	3063	73430
11	4176	0
12	4225	0

a problem due to lexical ambiguity which means that the same word can have multiple different meanings. The filter does not take this into account, so some news articles are mistakenly included, such as Air Jordans being interpreted as the country of Jordan and reuniting with someone as the Réunion, French island in the Indian sea. As most of the articles in the dataset are from 2023 and 2024, with 1,7% of the articles being from 2023 and 97,8% of the articles being from 2024, I focus my analysis on these two years alone. Together they form the majority 99,5% of the total dataset so they provide a comprehensive look into it.

Table 4.3: Label distribution in text classification dataset

yes	no
476	1524

As discussed with the publication date range and seen from the monthly distribution 4.2, data from November to December of 2024 is missing. To perform a complete analysis of one year worth of data, I use the data from November 2023 to October 2024.

4.2 Annotated data

From the Mission Grey News Dataset, I manually annotate a total of 3500 news articles for two tasks: text classification and sentiment analysis. For test classification tasks, 2500 articles are annotated by whether or they contain GPR themes or not. These are used for manually training, testing and evaluating the text classifier. For sentiment analysis tasks, 1000 articles are labeled into five distinct sentiment classes, depending on the overall tone of the article regarding GPR. These articles are used to fine-tune our sentiment classifier to make it better suited for this specific task.

4.2.1 Text classification

The first sample of 2000 articles from the year 2024 is taken from the Mission Grey News Dataset using the Pandas library for Python. These articles are manually annotated to either be considered geopolitical articles or not, marking them either as “1” for “yes” or “0” for “no”.

Topics I consider related to geopolitics contain information about conflicts, wars, trade wars, trade deals, natural resources and diplomacy between countries. Most of the included topics target two or more countries but I also consider topics targeting single countries, if the topic can affect the surrounding countries in the field of geopolitics, such as with conflicts and natural resources. Topics such as local politics, major accidents, travel guides and sports news are left out for not being relevant to our subject. This subset of data was split 80/20 for training and testing the text classification algorithm

Table 4.4: Example labels in the text classification annotated dataset

topic	label
Michael Jordan Manipulated and Crushed NBA Rival's Lifelong Desire, Admits Charles Barkley	0
Two Defendants Arrested for Conspiring to Illegally Export Weapons to South Sudan	0
Foreign movies filmed in Thailand generate up to 3 bn baht in income	0
Kazakhstan's Akhmet Ussen is on the golden pathway to take his second Asian title	0
Spain is the seventh happiest country in the world	0
Nadal gets even with De Minaur at Madrid Open but still doubts his body can hold up at French Open	0
Kenya flood toll rises to 181 as homes and roads are destroyed	0
Kiribati parliament votes to remove Australian judge	0
Fly2Sky Airlines partners with Air Serbia for the Summer season	0
Pinar de Rio's Vegueros leads alone in Cuban baseball championship	0
Higher education centers to focus on cooperation between Angola and Cuba	1
The Gambia, UK formally accept Agreement on Fisheries Subsidies; UK pledges to the Fund	1
Jordan slams 'extremist Israeli settlers' for dumping Gaza aid truck contents on street	1
Kazakhstan Says It's Ready to Host Azerbaijan-Armenia Talks; No Date Announced	1
Venezuela sends more troops to border with Guyana amid growing domestic challenges	1
Azerbaijan, Uzbekistan, and Kazakhstan will unite their energy systems	1
Multinational efforts launched to end devastating war in Sudan	1
Chairman of Cabinet of Ministers Akylbek Japarov to visit Turkmenistan	1
US Coast Guard boards Chinese fishing boats near Kiribati	1
Kuwait, Egypt urge Gaza truce	1

Table 4.5: Label distribution in sentiment analysis dataset

very negative	negative	neutral	positive	very positive
59	236	340	352	13

that extracts geopolitical news articles from the overall dataset.

A separate validation set of 500 articles is labelled in order to test the model on unseen data. This is done to better guarantee that we have differing data from the main 2000 labeled articles. Due to these 2000 random articles being taken by a random sample, I eliminate any possible overgeneralization should the topics in that random article sample be too similar to each other. Additionally as seen by table 4.3, the labels are not equally distributed. Train-test-splits for the 2000 article set were done randomly so in order to ensure there would be no tests on completely one-sided datasets, an additional validation set was utilized.

4.2.2 Sentiment analysis

For sentiment analysis, a second random sample of 1000 geopolitical classified articles is taken and annotated for fine-tuning. These articles were labeled as "0" for "very negative", "1" for "negative", "2" for "neutral", "3" "positive" or "4" "very positive".

In the "very negative" -category, I select articles with topics such as conflict and

Table 4.6: Example labels in the sentiment analysis annotated dataset

topic	label
A Slippery Slope? U.S., U.K. Launch Strikes on Iran-Backed Houthis	0
Mass evacuation in northern Iraq as Turkish forces cross border, mount incursion	0
Turkish Airstrikes Kill 17 Kurdish Militants in Northern Iraq	0
Armenia continues to deploy its troops along conditional border with Azerbaijan	0
Argentina calls for arrest of Iranian minister for suspected role in AMIA bombing	1
India blocks Bangladeshis fleeing chaotic regime change	1
Egypt expresses rejection to meddling with internal affairs of Sudan	1
Germany, Czech Republic accuse Russia of cyberattacks	1
New phase begins in Rail Baltica bridge construction over Neris in Lithuania	2
The two-way street of US-Slovak military ties	2
Roosevelt carrier arrives in South Korea in show of force to the North	2
Call to boost Cambodia, Malaysia economic ties	2
Kuwait Crown Prince Congratulates Pres. Biden On Independence Day	3
UNDP and Guinea-Bissau renew partnership to strengthen the response to HIV, Tuberculosis and Malaria in the Country	3
High-level Party meeting held between Vietnam and Laos	3
US elevates security relationship with Kenya at state visit	3
South Korea agrees to lend billions to Tanzania, Ethiopia	4
Serbia, Slovenia, Hungary agree to combine power exchanges	4
Sweden joins NATO's exercises as full member for the first time	4
Participants in Korea-Japan-China Business Summit Agree on Establishment of Working Group on Economic...	4

war. This category is reserved of the most GPR increasing topics. In the "negative" -category I select articles such as threats, tensions and trade wars. These also raise GPR but as much as those in the previous category. In the "neutral"-category I select topics that are overall neutral or if it is hard to tell whether or not it would have an effect on GPR. In the "positive"-category I select cooperative events, such as trade deals and general positive diplomacy between countries. In the "very positive" -category, I select events such as diplomatic deals spanning larger organizations or multiple countries, and providing support during war or conflict. This annotated data is used to fine-tune our sentiment analysis model for geopolitical news data labeling.

5 Method

In this section I propose an approach to construct a geopolitical risk index by classifying related news articles, extracting which countries the article discusses and calculating the overall sentiment. Classifying is done by first analyzing the accuracy of classification by a BERT-based model I fine-tune on my annotated geopolitical news data subset and keyword-based lexicon created by a LLM. Country extraction is done by using a state-of-the-art named entity recognition model. Sentiment analysis is done by a BERT-based model that I also fine-tune with another annotated geopolitical news dataset. These sentiments are stored in a dataframe, where the columns are countries found within the articles and rows indicate weeks, when the article was released. This enables us to build a comprehensive index with the possibility of displaying its weekly changes by a graph. Multiple different approaches to constructing indices from sentiment labels are explored and validated comparing them to the country-specific GPR index created by Caldara and Iacoviello [18].

5.1 Text Classification

For a geopolitical risk index I need to consider only news articles related to geopolitics. Two methods are explored here to filter them out from the Mission Grey News Dataset: text classification using distilBERT [38] and keyword-based lexicon by ChatGPT-o4-mini [39].

5.1.1 Classification using DistilBERT

BERT is chosen as a base approach to perform text classification due to its great general performance in text classification compared to other commonly used methods [40]. In its base model, it is too heavy for me to reasonably access and run, so DistilBERT [38] is chosen as a lighter version. The main task designed for it is either masked language modeling or next sentence prediction, but by adding a classification token to it, it can be used in text classification tasks.

I first need to fine-tune it on a down-stream task, our GPR news classification. Before actual training, I evaluate some hyperparameters to increase its performance. I explore three different ones: learning rate, batch sizes and epochs. Learning rate determines how much the model adjusts its parameters when it learns from new data. The ranges I test are 0.1, 0.01, 0.001, 0.0001, and 0.00001. Since I am using a pre-trained model, a smaller learning rate is sufficient since the model doesn't need to learn the whole task from zero. Batch size determines how many training samples are passed to the model at once. Larger batch sizes can result in faster and cheaper training but lower accuracy and smaller batch size can result in better accuracy but more computationally expensive training. I explored batch sizes of 16, 32, 64 and 128. Since the results can vary even with the general guidelines, testing different batch sizes for my current task is necessary to achieve great performance. Epochs control how many times the model is trained using the data. Passing the same data through the model multiple times gives the model ample time to learn and understand the patterns in it. However, passing it too many times can result in lower accuracy on unseen data and overfitting. I tested epochs of 2, 3, 4 and 5 for my hyperparameters. For my evaluation metric, I choose to use recall. This is due to a few factors. Firstly, false positives do not influence the analysis process with a significant factor. Even the news that are not related to geopolitics are concerned with local politics or sports results at first, so false positives are not that dangerous. Secondly, there are more non-geopolitical articles than geopolitical articles in the training set, the ratio is about 1-to-3. It is to be assumed that this kind of distribution is prevalent in the full dataset. Therefore, I want to capture all of the possible geopolitical occurrences due

to this skewed distribution. By favoring recall, our model becomes better at acquiring nearly all geopolitical data from unseen sources. Lastly, when we can prioritize recall over precision, the overall performance of the model can be improved over trying to balance both of them. Performing the hyperparameter selection gives diverse results. Most combinations result in a recall value around 0.6, some in a recall of 0, but one manages to reach a recall of 1. This might have been due to that model labeling all as positive, as testing that model yielded low values in other metrics. Due to this, the second best combination was chosen: learning rate of 0.001, batch size of 32 and 5 epochs.

5.1.2 Keyword-method using ChatGPT-o4-mini

Before deciding to use a text classifier to filter the news articles, I explored the possibilities of using a keyword-based search. Creating a comprehensive keyword lexicon requires a significant amount of domain knowledge, which I do not possess in the field of geopolitics. Large language models have seen increased performance during recent years, especially with zero-shot tasks, where detailed instructions are not provided [41]. I test the keyword-based method by prompting the GPT-o4-mini model to generate a keyword lexicon suitable for extracting geopolitical news articles. Detailed instructions are not given with the prompt, but the categories which I manually annotate the data were included to ensure the results could align with the text classification method. In addition to keywords, GPT-4o-mini provides short phrases related to geopolitical articles and some exclusion terms which help reduce the number of non-trivial news.

5.1.3 Classification results

Both methods are evaluated on the same validation set that they've not seen before to ensure equal validation and comparability of results. Even though I prioritized the recall metric with the text classifier, I decided to include all four metrics as a broad comparison between the two methods.

Both methods share similar metrics with the clear distinction being the significantly higher recall achieved by the distilBERT text classifier. This falls in line with small,

Table 5.1: Classification results

Method	Accuracy	Precision	Recall	F1-score
distilBERT	0.818	0.503	0.841	0.651
GPT lexicon	0.796	0.556	0.487	0.519

fine-tuned LLMs generally outperforming generative AI models in zero-shot tasks [42]. Since I am mostly interested in high recall, I perform the classification required for this study using the distilBERT text classifier.

5.2 Named entity recognition

For country-based indices NER is used to extract countries occurring in geopolitical news articles. Country detecting is a common task in NER so additional fine-tuning is not needed. Using the Huggingface library, I use the BERT-based fine-tuned named entity recognition model dslim/bert-base-NER [37] which achieves a f1-score of 91,7% on the CoNLL-2003 dataset [43]. This model is able to detect four distinct named entities: persons with tag: “PER”, organizations with tag: “ORG”, locations with tag: “LOC” and miscellaneous with tag: “MISC”. For named entities spanning multiple tokens/words, BIO-tagging is utilized. The named entity we need to capture countries with is the “LOC” tag. This raises three problems. Firstly, there are locations other than countries under the location tag. It cannot be utilized for country detection without some filtering. To combat this, I create a simple word list of all currently UN recognized countries. Each time a location entity is found, I check if the country list contains it. This can result in a lot of comparisons, but they are not computationally heavy enough to matter in this case. Secondly, some news articles might not have the countries mentioned in them but are still relevant. Such an article could talk about two prime ministers from different countries meeting. In these kinds of articles, they are referred to by their nationalities, which we need to capture. Luckily, the “MISC” tag detects nationalities from text. We can apply this the same way and we did for the countries: if the “MISC” tag entity is found in a separate nationality list, the corresponding country is fetched from the same

index in the country list. Thirdly, some countries have common abbreviations, some of which the NER algorithm cannot detect as entities. To alleviate this, a third filter to turn abbreviations into their common forms is created. In natural language processing, this is referred to as lemmatization. It's used more commonly to turn whole spans of text into their base forms, but since we are only interested in the countries, we benefit more from a custom filter rather than using existing lemmatization tools. For a quick and a comprhensive solution, I prompted GPT-4o-mini to list me most commonly used abbreviations for countries. These include cases such as "USA" referring to United States and "UAE" referring to United Arab Emirates.

5.3 Sentiment Analysis

Sentiment analysis is used to classify news articles into either heightening geopolitical risk or lowering it. The news article text is passed to the model, which then labels it between very negative, implying higher geopolitical risk, and very positive, implying lower geopolitical risk. These labels are stored in an array with five values, each representing a different label. These arrays in turn are stored in the main dataframe, where columns represent different countries and rows the different weeks articles are published on. In order to locate correct arrays to update, the columns are compared to the countries extracted from the news article with NER and the rows compared to the article's publication date's week. Then, the correct array(s) are updated by increasing the values of the corresponding label occurrences.

5.4 The index

There are a multitude of approaches to transform sentiment ratios into indices. I explore two different methods established during earlier studies. Burns's method [24] calculates the total sum of the sentiment. Caldara and Iacoviello's [18] method involves constructing a ratio by comparing geopolitical risk increasing news to the total number of news articles. I evaluate this approach twice: by comparing the country's articles to its own article total

and to the total number of articles for that week from all countries. As a third method, my own method of taking the "ratio of news" -method, but amplifying the effects of "very negative" -events to contribute twice as much as the "negative" events to the overall index, is also explored as a possible way. Indices created by these three methods are then compared to the existing GPR index created by Caldara and Iacoviello in order to evaluate them.

5.4.1 Sum of sentiment

In his case study, Burns [24] explores the sum of sentiment from Twitter/X post data to study the effects of the Ukraine War on financial markets. He constructs a time series based on these sentiments to highlight how total sentiment from Twitter/X changes when the war starts. Similarly to his approach, I use the sum of sentiment to construct an index revolving around sentiment analysis events. Each of the five labels is given a value from -2 to 2, with very positive events labeled as -2, positive events as -1, neutral events as 0, negative events as 1 and very negative events 2. These can be combined into a sum of sentiment formula 5.1.

Sum of Sentiment:

$$(vneg * 2) + neg - pos + (vpos * 2) + 100 \tag{5.1}$$

The total sum of these event occurrences gives us the total overall sum of sentiment for a specific week and country. This can be added to a set value such as 100, to create a default base value for the index. Values over 100 indicate higher geopolitical risk and values lower than 100 indicate lower geopolitical risk. We can combine these weekly country sums to construct time series indices for each country found within the news data.

At the start of the year, there was less news overall which is reflected by the low changes at the start. There are major fluctuations towards both high and low geopolitical risk in the middle. I think this index does not work well when you have different counts of

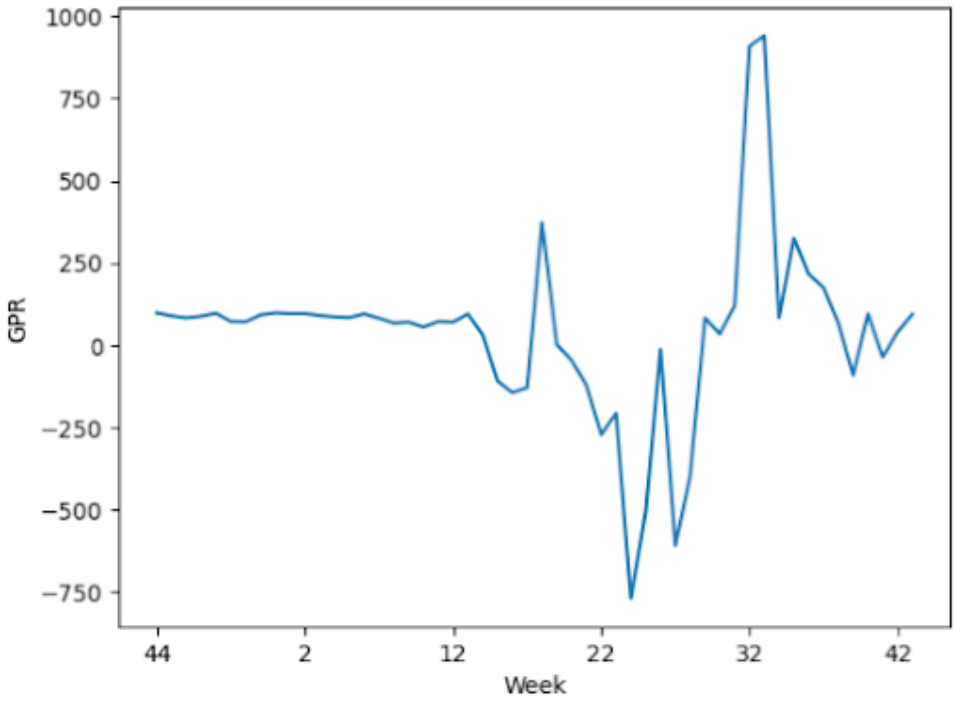

Figure 5.1: Sum of sentiment GPR index for Russia 2023-2024

articles for each week. It also reacts strongly should there be significantly more positive and negative news for a single week.

5.4.2 GPR news ratio

In their paper, Caldara and Iacoviello [18] use a method comparing the number of geopolitical risk-increasing news to the number of total news. I utilize a method similar to this by comparing the number of country's news articles to that country's total news articles for a specific week. This yields a ratio of news. This ratio can be multiplied by a desired value, such as 100, to achieve a comparable index. They used a binary classification method to either mark the article as relevant from the viewpoint of geopolitical risk or not. For this reason, I treat the "negative" and "very negative" labels similarly. It's to be noted that Caldara and Iacoviello do not use "very negative" or "very positive" which I will explore later. I construct the GPR news ratio by calculating the total number of "very negative" and "negative" news articles, then dividing that by the total number of all five categories of news articles and multiplying the result by 100. To ensure better comparison with the earlier method, I add the resulting value to 100. This produces the

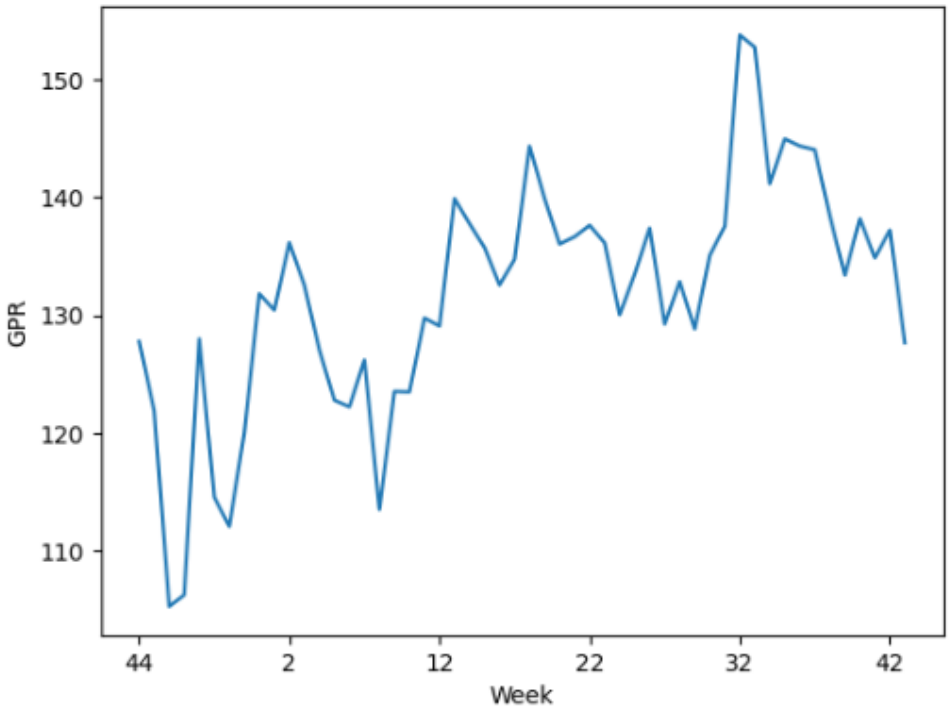

Figure 5.2: Country-based news ratio GPR index for Russia 2023-2024

country-based news ratio formula 5.2.

Country-based news ratio:

$$(100 * \frac{vneg + neg}{vneg + neg + neutral + pos + vpos}) + 100 \tag{5.2}$$

Using this formula, we can calculate weekly geopolitical risk totals for each country. These can be then combined into a time series to highlight changes and for easier comparison.

Unlike sum of sentiment, GPR news ratio method adjusts itself to the number of articles for that week. It fluctuates less and provides more articulate information about that week's GPR. The overall trend indicates growing geopolitical risk over the year.

For country specific indices, Caldara and Iacoviello only report the ratio for a country's geopolitical risk increasing news compared to the total number of news for that week. In order to compare my overall pipeline approach to theirs, I use the total news ratio formula 5.3 which gives me the weekly ratio for a single country. These weekly ratios can then be combined into a weekly time series. The overall changes in this index seem to

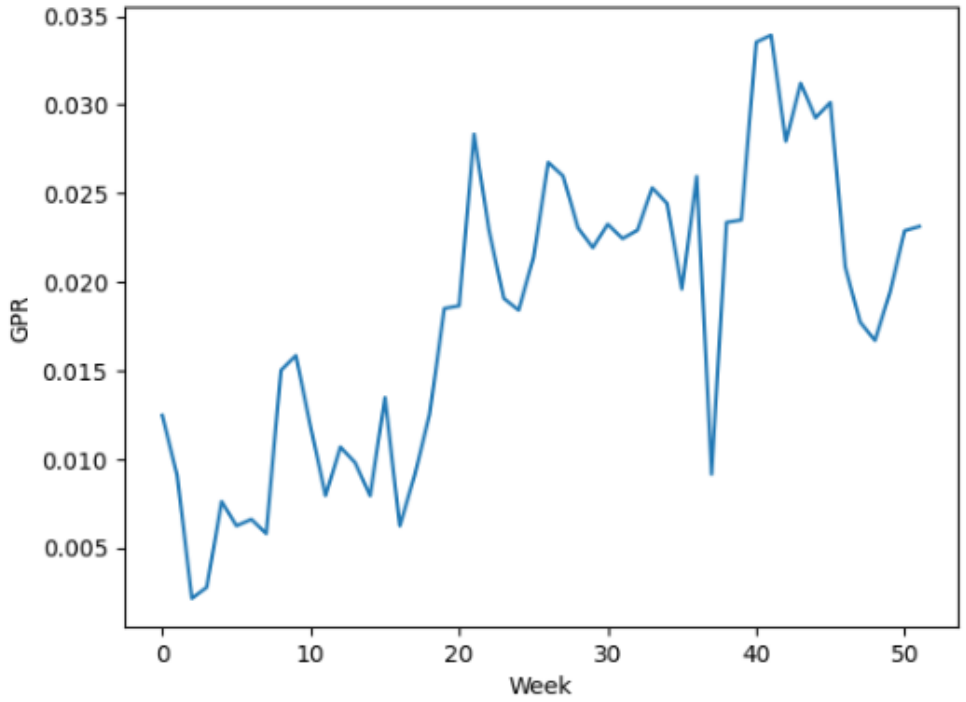

Figure 5.3: Total news ratio GPR index for Russia 2023-2024

mimic those of that in the country based news ratio index.

Total news ratio:

$$\frac{vneg + neg}{sum(articles)} \tag{5.3}$$

5.4.3 Weighted GPR news ratio

My sentiment analysis method produced additional labels, "very negative" and "very positive" compared to what Caldara and Iacoviello used. I want to better highlight these extreme labels by weighting them. For simplicity, I choose to take an approach similar to the sum of sentiment by weighting them to be twice as important than their regular versions. I take these weights into account when constructing the GPR news ratio index. This forms the weighted news ratio formula 5.4.

Weighted news ratio:

$$(100 * \frac{(2 * vneg) + neg}{(2 * vneg) + neg + neutral + pos + (2 * vpos)}) + 100 \tag{5.4}$$

The method is similar to GPR news ratio, but "very negative" and "very positive"

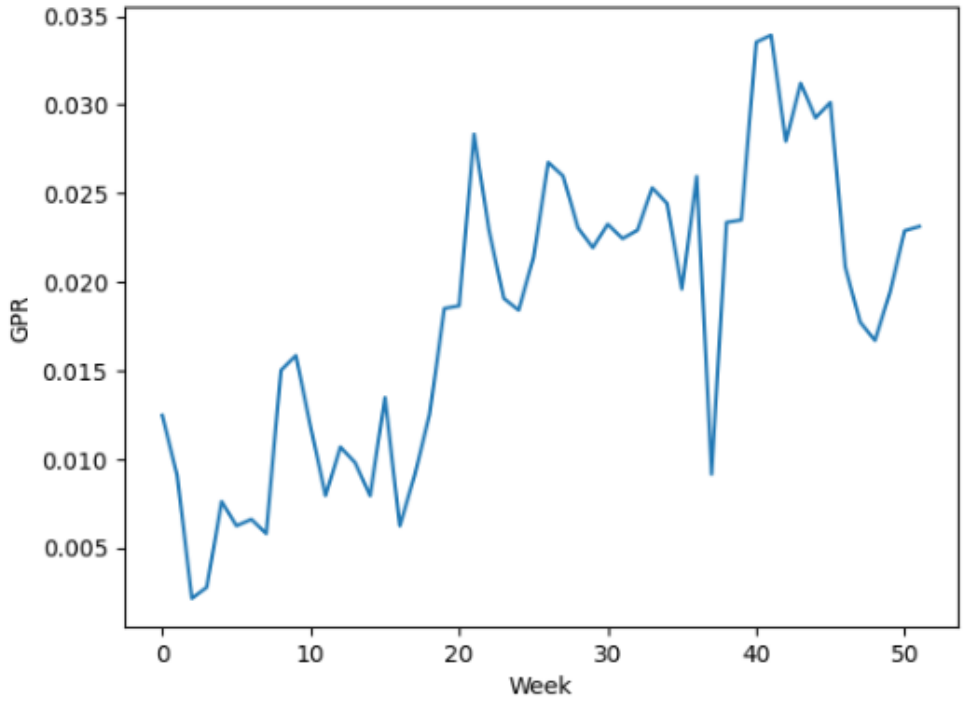

Figure 5.4: Weighted news ratio GPR index for Russia 2023-2024

labels amounts are multiplied by two before the calculation. By calculating these for every week in our data, we can construct a time series index for a single country index.

The weighted GPR news ratio index looks to be similarly distributed to the other news ratio methods. The added weights do not seem to affect the overall changes in the index very much.

5.5 Evaluation

The indices are evaluated through comparing them to each other. For each of them I modify the GPR index of Russia from November 2023 to October 2024 since due to the ongoing war with Ukraine and tensions with NATO, geopolitical risk should be prominently visible in their GPR indices. Due to the sum of sentiment receiving greatly fluctuating values compared to the other indices, it is not included as part of the comparison since it undermines the changes seen in the others if plotted together. I first compare the country news ratio to the weighted news ratio method to see how the weighting affected the index. From the results of that, I choose the better index to be compared secondly to the total news ratio method. The total news ratio method must be upscaled slightly

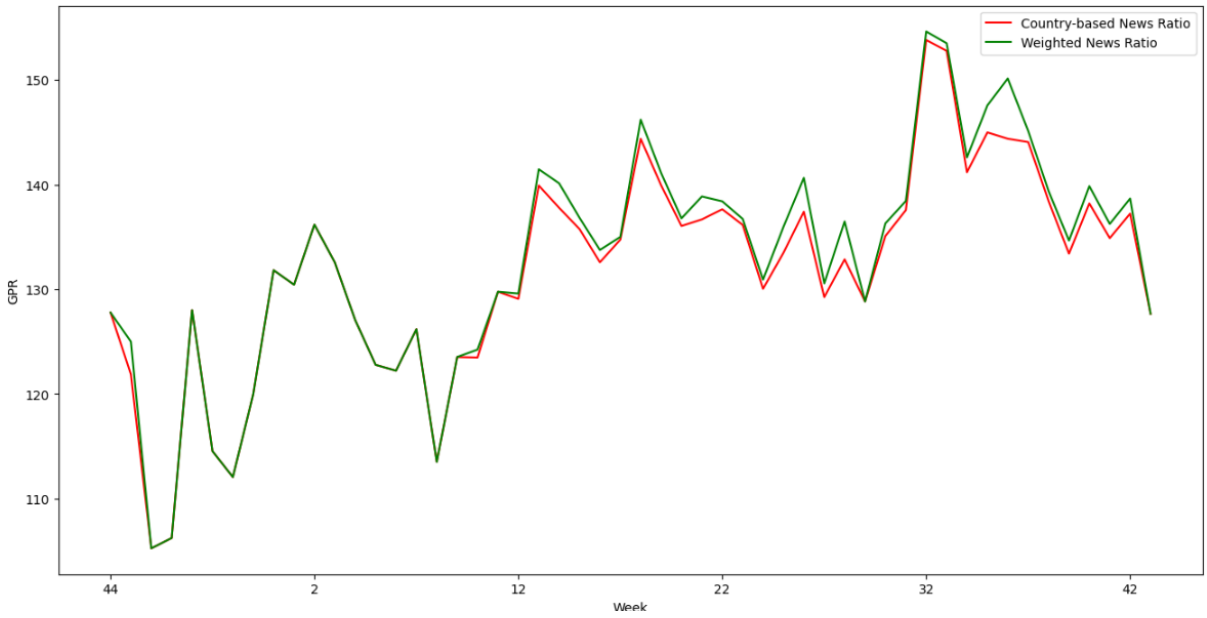

Figure 5.5: Comparison of country-based and weighted news ratio indices

to enable comparison to the other methods. Thirdly, the total news ratio is compared to the already evaluated Caldara and Iacoviello GPR index. An overall summary and general thoughts on the evaluation are discussed at the end.

Table 5.2: Sentiment transformation methods

Index	**Formula**
Sum of sentiment	$(2 \cdot vneg) + neg - pos + (2 \cdot vpos) + 100$
Country-based news ratio	$100 \cdot \frac{vneg + neg}{\sum(\text{labels}))} + 100$
Total news ratio	$\frac{vneg + neg}{\sum(\text{articles})}$
Weighted news ratio	$100 \cdot \frac{2 \cdot vneg + neg}{2 \cdot vneg + neg + neutral + pos + 2 \cdot vpos} + 100$

Comparing the effects of weighting the labels, there does not seem to be much change to the overall index. When there is a peak in geopolitical risk, the weighted news ratio reacts a bit stronger to it. This is especially visible when there's a flatter peak from the country-based method. These indices do not seem to go under a 100, so they cannot be used to quantify lower geopolitical risk as efficiently as the sum of sentiment method. I could argue that if we want to highlight the changes in geopolitical risk better, the weighted news ratio would be a better choice overall.

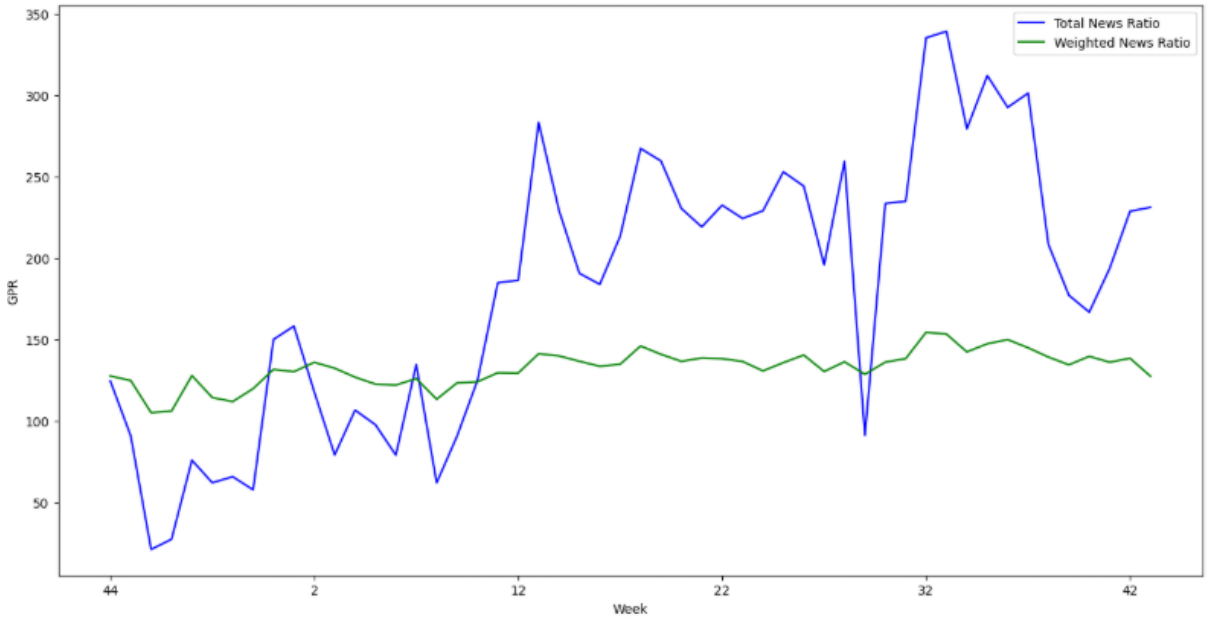

Figure 5.6: Comparison of total and weighted news ratio indices

Comparing the total news to weighted news ratio could not be done directly. In order to move them to a similar scale, I multiply the total news ratio by 10000. The total news ratio fluctuates more compared to the weighted news ratio. Even though there are more clear changes, they both change similarly. Both peak and drop in the same spots, even if the weighted ones are more clearly noticeable. I could argue that the total news ratio method is better at communicating the overall geopolitical state of the world due to the comparisons to all countries. The more visible changes make it easier to tell how geopolitical risk has changed, especially with some major events.

Comparing my best index to Caldara and Iacoviello's GPR index shows that while their index suggests a higher geopolitical risk, our indices follow nearly the same shape. They use either monthly or daily data, so I take the weekly averages of my index and construct a monthly time series from it. The differences in GPR can be explained by many factors but the main contributor is that they used a different data set. While the distribution of my dataset should not directly affect the overall index, the increased amounts on some weeks might be due to excessive coverage of war topics, which would in turn have a visible effect on the index. Depending on what newspapers they use, the geopolitical topics might have been overall covered on a different scale. The most

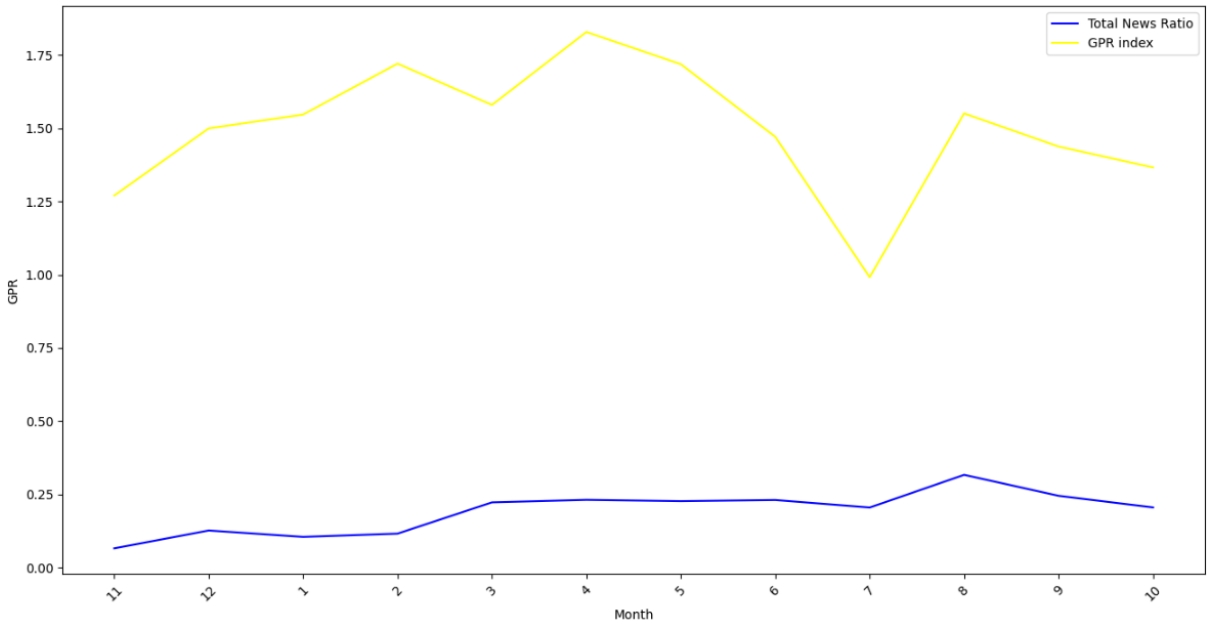

Figure 5.7: Comparison of total news ratio and Caldara&Iacoviello GPR index

important thing that I could take from this comparison is the overall similarity of shape between these two indices. At the start, there's a gradual increase in geopolitical risk with both of them. There's also a decline in July and a peak in August in both of them. This suggests that while I might not have the same data, the overall trends in geopolitical risk are similar. To better perform a comprehensive evaluation, dates on some larger GPR events from the overall timeframe these indices present could be utilized here. It would give some background whether or not these indices react to it and at what scale.

In summary, after comparing these different approaches and looking at them overall they all seem to indicate some degree of GPR altogether. Sum of sentiment seems to suffer from the overall distribution of the news articles but this does not seem to affect the other methods. However, sum of sentiment seems to highlight the aspect of lowered geopolitical risk much better than the other methods. If this is a desirable metric that also needs to be clearly visible, the sum of sentiment method seems to be a great option. For more general and better real-time GPR index modeling I would use any of the ratio methods explored. This is especially important if we have a real-time index utilizing constant, new news data. There's no explicit guarantee that every day, week or month has the same number of articles which could cause skewness with the sum of sentiment method. As

seen with the ratio methods, since the ratio does not necessarily suffer whether or not there's several or few news articles, it would work well with varying amounts of data. The ratio methods seem to share the same peaks and lows but due to the total news ratio having a different scale of values, it needs to be multiplied by a factor in order for it to be comparable to the others. It was comparable to the existing GPR index by Caldara and Iacoviello [18] and performed the best overall due to this.

6 Results

In this thesis I constructed a comprehensive pipeline to extract GPR indices from news data 6.1. Through literature review of the field, key points of idenftifying GPR news and countries were explored by natural language processing methods. Evaluation of BERT-based encoder and GPT-based keyword approach showed that even with a small fine-tuning to task data, encoders are a great choice to label geopolitical news articles. Utilizing NER instead of traditional keyword extraction is a quick and efficient way to implement country data extraction from news articles. For the GPR labeling process, sentiment analysis proved to be a flexible method that can be fine-tuned for new terminology fairly easily. It could be used to produce multiple different kinds of GPR indices.

To sum up the answer for the research question 1, five components are needed to acquire GPR indices through news data. You need a diverse dataset, a way to classify GPR articles, a way to extract countries from these articles, a way to label the articles by level of GPR and a formula to turn these labels into GPR indices. Utilizing NLP methodologies alongside these components can result in a pipeline capable of processing large quantities of data in real time. As for research question 2, there are multiple ways to transform sentiment into indices. I used two overall categories of summing up the amount of positive and negative sentiment found within the articles or comparing the frequency of negative articles in all of the articles. Each of these had some benefit or drawback to them but the frequency comparison seems to be a good general way to use as an indice transformation method.

Resulting indices all modeled geopolitical risk to some degree. Methods like sum of sentiment seemed to suffer from skewed distribution but managed to highlight the as-

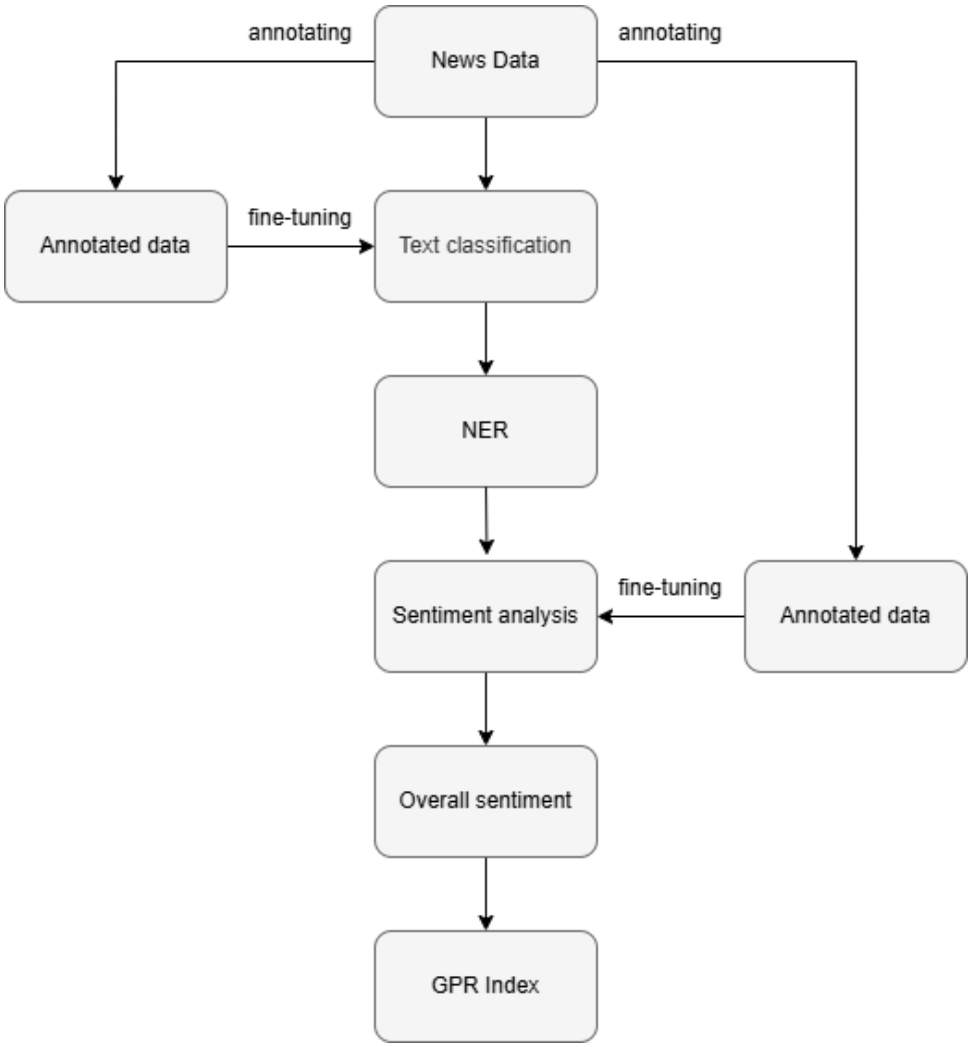

Figure 6.1: Pipeline approach

pect of lowered geopolitical risk better than the other methods. Different ratio methods explored did not care about distribution or number of news articles, each of them managing to capture similar peaks and lows in the overall GPR. Weighting some sentiment labels did not affect the results much. This effect could be compared better by acquiring more data with weightable labels. Total news ratio showed the best overall results, being similar in changes to Caldara and Iacoviello's GPR index [18].

Constructing the approach in this thesis could have benefited from some quality improvements. Having overall more resources and time available, the text classification algorithm could have been more accurate. Due to the models being heavy to run, even with 40 gigabytes of VRAM, I had to choose a lighter one which might have affected the results. I had only 2500 articles of training data due to having to manually annotate it. I could not find exact training data for this type of domain, so more annotated data or a model specifically trained for this type of task could have improved the overall accuracy metrics of my classification. Testing different hyperparameters also increases exponentially in time when introducing a new variable to the mix, so I could only explore some basic ones.

7 Future work

My thesis mainly focused on identifying GPR, but the general pipeline does not necessarily have to model just it. This kind of approach could be done for other country specific measurable qualifications, such as general left/right-leaniness in politics or environmentalism. Utilizing a text classifier and sentiment classifier trained on different kinds of data, the pipeline could possibly be generalized to work with other topics. I think the main limiter here is the sentiment analysis model and what kind of sentiment it can extract from the text but especially with something that has either a positive or negative quality to it this pipeline could work as a basis for a future research topic.

The news dataset used for this method was completely in English. Multilinguality is argued to be a vital aspect when performing this kind of analysis. "Carrying out sentiment analysis in only one language increases the risk of missing essential information authored elsewhere." [44]. Multilinguality has already been explored when constructing indices [24] but how it could be implemented into a pipeline like this could be a complex and an interesting topic to research more about.

To transform sentiment into indices, I use pretty basic methods such as sum and frequency. Some simple weighting systems for certain labels were also explored. This part of the pipeline alone could be a topic of its own. I could not find many different kinds of methods available to perform this so an exploratory study to just focus on this aspect alone could be conducted.

In the field of NLP, other methods could also be utilized as a part of the pipeline approach. Topic modeling is a field that aims to extract topic keywords from a set of documents. It could be used to validate sets of geopolitical risk articles and even classify

them based on trade wars, conflicts and cybersecurity threats. In addition to validation, the overall classification accuracy could be improved by using multiple text classifiers in a hybrid manner.

References

[1] B. Eichengreen, "Geopolitics and the global economy", *Journal of International Money and Finance*, vol. 146, p. 103 124, 2024.

[2] X. Wang, Y. Wu, and W. Xu, "Geopolitical risk and investment", *Journal of Money, Credit and Banking*, vol. 56, no. 8, pp. 2023–2059, 2024.

[3] A. Rose, Z. Chen, and D. Wei, "The economic impacts of russia–ukraine war export disruptions of grain commodities", *Applied Economic Perspectives and Policy*, vol. 45, no. 2, pp. 645–665, 2023.

[4] J. W. Glauber, D. Laborde, and A. Mamun, "From bad to worse: How russia-ukraine war-related export restrictions exacerbate global food insecurity", *IFPRI book chapters*, pp. 92–96, 2023.

[5] E. Rusanti, A. F. Isman, N. Nashrullah, A. Mansyur, and A. A. Elzaanin, "Israel-palestine conflict: Tracking global economic responses and fears", *Shirkah: Journal of Economics and Business*, vol. 10, no. 1, pp. 1–19, 2025.

[6] Z. Li, P. Farmanesh, D. Kirikkaleli, and R. Itani, "A comparative analysis of covid-19 and global financial crises: Evidence from us economy", *Economic Research-Ekonomska Istraživanja*, vol. 35, no. 1, pp. 2427–2441, 2022.

[7] C. A. McClelland and G. D. Hoggard, *Conflict patterns in the interactions among nations*. University of Southern California, 1968.

[8] M. D. Ward, "Seasonality, reaction, expectation, adaptation, and memory in cooperative and conflictual foreign policy behavior: A research note", *International Interactions*, vol. 8, no. 3, pp. 229–245, 1981.

[9] D. B. Bobrow, "Uncoordinated giants", *Foreign Policy USA/USSR. Beverly Hills: Sage*, 1982.

[10] W. J. Dixon, "Measuring interstate affect", *American Journal of Political Science*, pp. 828–851, 1983.

[11] J. S. Goldstein and J. R. Freeman, *Three-way street: Strategic reciprocity in world politics*. University of Chicago Press, 1990.

[12] J. S. Goldstein, "A conflict-cooperation scale for weis events data", *Journal of Conflict Resolution*, vol. 36, no. 2, pp. 369–385, 1992.

[13] J. E. Vincent, "National attributes as predictors of delegate attitudes at the united nations1", *American Political Science Review*, vol. 62, no. 3, pp. 916–931, 1968.

[14] Y. Feng, "Measuring international conflict: Developing cross-country time-series data", *International Interactions*, vol. 26, no. 3, pp. 287–319, 2000.

[15] E. E. Azar, "The conflict and peace data bank (copdab) project", *Journal of Conflict Resolution*, vol. 24, no. 1, pp. 143–152, 1980.

[16] P. A. Schrodt, S. G. Davis, and J. L. Weddle, "Political science: Keds—a program for the machine coding of event data", *Social Science Computer Review*, vol. 12, no. 4, pp. 561–587, 1994.

[17] P. T. Brandt, J. R. Freeman, and P. A. Schrodt, "Real time, time series forecasting of inter-and intra-state political conflict", *Conflict Management and Peace Science*, vol. 28, no. 1, pp. 41–64, 2011.

[18] D. Caldara and M. Iacoviello, "Measuring geopolitical risk", *American economic review*, vol. 112, no. 4, pp. 1194–1225, 2022.

[19] J. Klement, *Geo-economics: The interplay between geopolitics, economics, and investments*. CFA Institute Research Foundation, 2021.

[20] S. Kapadia, *Evaluate topic models: Latent dirichlet allocation (lda)*, `https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0`, Accessed: 2021-03-09, 2021.

[21] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis", *Knowledge-Based Systems*, vol. 69, pp. 24–33, 2014.

[22] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text", in *Proceedings of the international AAAI conference on web and social media*, vol. 8, 2014, pp. 216–225.

[23] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in arabic pre-trained language models", *arXiv preprint arXiv:2103.06678*, 2021.

[24] J. C. Burns, "Automatic evaluation of geopolitical risk", Ph.D. dissertation, The University of St Andrews, 2024.

[25] Q. Li, H. Peng, J. Li, *et al.*, "A survey on text classification: From traditional to deep learning", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–41, 2022.

[26] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[27] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", in *European conference on machine learning*, Springer, 1998, pp. 137–142.

[28] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks", *arXiv preprint arXiv:1503.00075*, 2015.

[29] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences", *arXiv preprint arXiv:1404.2188*, 2014.

[30] J. Chen, Z. Gong, and W. Liu, "A nonparametric model for online topic discovery with word embeddings", *Information Sciences*, vol. 504, pp. 32–47, 2019.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[32] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, *Improving language understanding by generative pre-training.(2018)*, 2018.

[33] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 347–354.

[34] S. Taj, B. B. Shaikh, and A. F. Meghji, "Sentiment analysis of news articles: A lexicon based approach", in *2019 2nd international conference on computing, mathematics and engineering technologies (iCoMET)*, IEEE, 2019, pp. 1–5.

[35] tabularisai, *tabularisai/robust-sentiment-analysis*, `https://huggingface.co/tabularisai/robust-sentiment-analysis`, A DistilBERT-based sentiment classification model fine-tuned on synthetic data (5 classes)., 2025.

[36] A. Akhundov, D. Trautmann, and G. Groh, "Sequence labeling: A practical approach", *arXiv preprint arXiv:1808.03926*, 2018.

[37] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding", *CoRR*, vol. abs/1810.04805, 2018. arXiv: `1810.04805`. [Online]. Available: `http://arxiv.org/abs/1810.04805`.

[38] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter", *arXiv preprint arXiv:1910.01108*, 2019.

[39] OpenAI, "Gpt-4o system card", OpenAI, Technical Report, 2024, Omnimodal model capabilities, limitations, safety evaluations, and Preparedness Framework overview. [Online]. Available: `https://openai.com/index/gpt-4o-system-card/`.

[40] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing bert against traditional machine learning text classification", *arXiv preprint arXiv:2005.13012*, 2020.

[41] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners", *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[42] M. J. J. Bucher and M. Martini, "Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification", *arXiv preprint arXiv:2406.08660*, 2024.

[43] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition", *arXiv preprint cs/0306050*, 2003.

[44] I. Karageorgou, P. Liakos, and A. Delis, "A sentiment analysis service platform for streamed multilingual tweets", in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2020, pp. 3262–3271.