Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR WIRTSCHAFTS- UND
SOZIALWISSENSCHAFTEN

Identifizierung von Dekarbonisierungsversprechen von Firmen mittels Large-Language-Modellen

BACHELORARBEIT

zur Erlangung des Bachelorgrades
an der Fakultät für Wirtschafts- und Sozialwissenschaften
Fachbereich Volkswirtschaftslehre
der Universität Hamburg

Erstgutachter: Prof. Dr. Ole Wilms
Prof. für VWL, insb. Makroökonomik und Fiskalpolitik

Autor: **Marcel Weschke**

Studiengang: B.Sc. Volkswirtschaftslehre

Abgabedatum: 20. Januar 2025

Zusammenfassung

Marcel Weschke

Thema der Bachelorarbbeit
Identifizierung von Dekarbonisierungsversprechen von Firmen mittels Large-Language-Modellen

Stichworte
Universität Hamburg, Volkswirtschaftslehre, Bachelorarbeit, Large-Language-Modelle, LLM, Klassifizierung, Tokenisierung, GPT-4, Gemma 2, Llama 3.1, ClimateBERT Net-Zero, USA, Nachrichtenartikel, CO_2-Emissionen, Grüne Versprechen, Genaunigkeit, Präzision, Richtig-positiv-Rate, F-Maß, Richtig-negativ-Rate, Konfusionsmatrix,

Kurzzusammenfassung
Diese Bachelorarbeit untersucht den Einsatz von großen Sprachmodellen (im Englischen: Large Language Models, kurz LLMs), im Kontext der Ökonomie und Finanzwirtschaft, mit einem speziellen Fokus auf die Textklassifikation. Der Schwerpunkt liegt dabei auf der Analyse von Nachrichtenartikeln, die freiwillige Unternehmensankündigungen zu geplanten Reduzierungen von CO_2-Emissionen behandeln. Verschiedene LLMs werden basierend auf ihrer semantisches Verständnis und ihre Klassifikationsleistung verglichen, um Gemeinsamkeiten und Unterschiede herauszuarbeiten.

Inhaltsverzeichnis

Abbildungsverzeichnis iv

Tabellenverzeichnis iv

Formelzeichen und Abkürzungen v

1 Einleitung 1
1.1 Ziele und Problemstellung 2
1.2 Methodik 3

2 Theoretische Grundlagen 5
2.1 Maschinelles Lernen 5
2.1.1 Einsatz von LLMs in der Ökonomie und Finanzwirtschaft . 5
2.1.2 Lernstile des maschinellen Lernens 7
2.2 Text Mining / Computerlinguistik 8
2.3 Transformermodelle 9

3 Modellentwicklung und -Implementierung 11
3.1 Datengrundlage 11
3.2 Datenanalyse 11
3.3 Data Mining 12
3.4 Datenvorverarbeitung 12
3.5 Embedding 15
3.6 Textklassifikationsverfahren 16
3.7 Bewertung von Klassifikationen 17

4 Modellanalyse und -Ergebnisse 19
4.1 Aufbau 20
4.2 Experimente 21
4.3 Auswertung 26

5 Schlussbetrachtung 33
5.1 Zusammenfassung 33
5.2 Ausblick 33

Literatur 35

Anhang 37

Abbildungsverzeichnis

1.1 Auszug des Kernbereichs der LLM Gesamtübersicht 4
3.1 Exemplarischer GPT-4o Tokenizer Prozess 13
3.2 Exemplarischer Embeddingprozess 15
3.3 Übersicht der Klassen-Klassifikationstypen 16
4.1 Schematischer Prozessablauf der LLM-Analyse 20
4.2 Prozentuale Verteilung positiver Artikel-Klassifikationen (*GPT-4*) 21
4.3 Prozentuale Verteilung positiver Artikel-Klassifikationen (*Gemma 2*) 23
4.4 Prozentuale Verteilung positiver Artikel-Klassifikationen (*Llama 3.1*) 24
4.5 Prozentuale Verteilung positiver Artikel-Klassifikationen (*Climate-BERT-NetZero*) 25
4.6 Übersicht der GPT-4-Modellergebnisse in Form einer Konfusionsmatrix mit Metriken. 29
A.1 Umfassende Übersicht von LLMs 38
A.2 Trend der über die Jahre veröffentlichten wissenschaftlichen Artikel mit direktbezug zu LLM spezifischen Schlüsselwörtern 39
A.3 Chronologische Darstellung der LLM-Veröffentlichungen 39
A.4 Exemplarisches Klassifikationsbeispiel: LLM (Llama 3.1) & Experte 40
A.5 Grafische Gesamtübersicht der Modellergebnisse 41
A.6 Übersicht der GPT-4-Modellergebnisse in Form einer Konfusionsmatrix mit Metriken. 42
A.7 Übersicht der Gemma-2-Modellergebnisse in Form einer Konfusionsmatrix mit Metriken. 42
A.8 Übersicht der Llama-3.1-Modellergebnisse in Form einer Konfusionsmatrix mit Metriken. 42

Tabellenverzeichnis

4.1 Ergebnisübersicht verwendeter LLMs für den Klassifizierungsprozess 28
4.2 Ergebnisübersicht der LLM-Klassifizierungsmetriken 31

Formelzeichen und Abkürzungen

Symbole

i		Iterator $i = 1, 2, ..., N$
L		Lösungsmenge der Klassifikationsklassen
m		Steigungsparameter einer linearen Funktion
N		Gesamtmenge der Nachrichtenartikel
t	J	Zeit in Jahren
x_i		Vorhergesagte Klasse
y_i		Richtige Klasse
$x\,;\,y$		Klassifikationsstichprobe
$\overline{x}\,;\,\overline{y}$		mittlere Stichprobenmengen
f_n		Anzahl an fälschlicherweise als negativ eingestuften Klassen
f_p		Anzahl an fälschlicherweise als positiv eingestuften Klassen
r_n		Anzahl an richtigerweise als negativ eingestuften Klassen
r_p		Anzahl an richtigerweise als positiv eingestuften Klassen

Abkürzungen

CO_2 Kohlenstoffdioxid 1, 19, 34

COP21 21. UN-Klimakonferenz in Paris 1

GPT Generativer vortrainierter Transformer 2–4, 7, 14, 16, 19, 21, 22, 25–31, 33

IRA Inflation Reduction Act 1

KI Künstliche Intelligenz 5, 19

KNN Künstliche neuronale Netze 17

kNN k-Nächste-Nachbarn-Klassifikation 10

LLMs große Sprachmodelle 1–8, 12, 13, 16, 20–22, 24, 28–30, 33

RAM Arbeitsspeicher 2

RF Random Forest 10

SVM Support-Vektor-Maschine 10

THG Treibhausgasemissionen 1, 3

USA Vereinigte Staaten von Amerika 1, 26

1 Einleitung

Die Untersuchung freiwilliger Unternehmensverpflichtungen zur Dekarbonisierung wird angesichts der wachsenden Bedeutung nachhaltiger Unternehmensstrategien und der globalen Notwendigkeit, den Klimawandel zu bekämpfen, in naturwissenschaftlichen und wirtschaftlichen Kontexten immer relevanter. In dieser Arbeit wird der Einsatz von großen Sprachmodellen (im Englischen: Large Language Models, kurz LLMs zur Analyse von US-amerikanischen Wirtschaftsartikeln im Zusammenhang mit Marktbewegungen und freiwilligen Unternehmensverpflichtungen untersucht, insbesondere in Bezug auf Dekarbonisierungsziele. Die Bedeutung der Vereinigten Staaten von Amerika (im Englischen: United States of America, kurz USA) im globalen Klimakontext resultiert aus ihrer Rolle als einer der größten Treibhausgasemittenten und ihrer Einflussnahme auf die internationale Klimapolitik. Als der zweitgrößte Emittent von Kohlenstoffdioxid (CO_2), der im Jahr 2018 für etwa 15% der globalen Emissionen verantwortlich war (vgl. Ahmad u. a. (2023); gemäß Anhang A2) leisten die USA einen wesentlichen Beitrag zur globalen Emissionslast. Im Jahr 2007 haben die USA ihre historischen Höchstwerte bei den Emissionen verzeichnet. Bis 2020 gelang es, diese um 21% zu senken – zum Teil aufgrund wirtschaftlicher und pandemiebedingter Einflüsse. Das Pariser Abkommen von 2015 setzte einen entscheidenden Rahmen für Klimaziele, dem sich die USA mit dem Ziel angeschlossen haben, ihre Emissionen bis 2030 um 50 – 52% im Vergleich zu 2005 zu senken. Dieses auf der 21. jährlichen Klimakonferenz der Vereinten Nationen (COP21) beschlossene Abkommen, stellte einen Wendepunkt in den Klimaverhandlungen dar, da sich nahezu 200 Nationen verpflichteten, bis 2050 Netto-Null-Treibhausgasemissionen (im Englischen: Net-Zero emissions) zu erreichen (vgl. Acharya, Engle und Wang (2025)). Hierbei bezeichnet Net-Zero das Gleichgewicht zwischen der Menge an ausgestoßenen Treibhausgasen (THG) und der Menge, die aus der Atmosphäre entfernt wird. Dieses Gleichgewicht kann durch eine Kombination aus Emissionsreduktionen und Emissionsentfernung erreicht werden und stellt somit eine zentrale Zielsetzung für die globale Klimapolitik dar. Der *Inflation Reduction Act (IRA)* von 2022 markiert einen weiteren zentralen Schritt, da er umfangreiche Investitionen in erneuerbare Energien und innovative Technologien ermöglicht.

Die Betrachtung der USA bietet daher eine wichtige Grundlage, um wirtschaftliche und politische Zusammenhänge im globalen Klimaschutz zu analysieren und deren weitreichende Auswirkungen zu verstehen. LLMs bieten eine innovative Methode, große Mengen an Textdaten effizient zu verarbeiten und Sprachbarrieren weitestgehend zu überwinden. LLMs sind bereits heutzutage in der Lage, Texte aus einer Vielzahl von Sprachen zu analysieren, wodurch Unternehmen eine globale Perspektive auf Markt- und Unternehmensdynamiken gewinnen können.

Der wesentliche Neuheitsaspekt dieser Technologie liegt in ihrer Fähigkeit, den Bedarf an Personalkosten erheblich zu reduzieren, wodurch sowohl zeitliche als auch finanzielle Ressourcen optimiert werden. Dies führt zu einer Entlastung der Entscheidungsträger und ermöglicht es ihnen, sich intensiver mit der Auswertung der Ergebnisse und der Implementierung geeigneter Strategien zu befassen. Dadurch können sie schneller auf Veränderungen am Markt reagieren.

Im Mittelpunkt stehen die folgenden grundlegenden Fragen:

[1] Wie effektiv sind die gewählten LLMs beim Auffinden neuer und freiwilliger Dekarbonisierungsverpflichtungen mittels automatisierter Klassifikationen von US-amerikanischen Nachrichtenartikeln?

[2] Welche Unterschiede, in der Performance, ergeben sich bei der Anwendung verschiedener LLMs, zur Sentimentanalyse von neuen und freiwilligen Dekarbonisierungsverpflichtungen in US-amerikanischen Nachrichtenartikeln?

[3] Inwieweit können die verschiedenen LLMs den semantischen Kontext der neuen und freiwilligen Dekarbonisierungsverpflichtungen in den US-amerikanischen Nachrichtenartikeln identifizieren und klassifizieren?

1.1 Ziele und Problemstellung

Das Ziel wird darin bestehen, zu ermitteln, wie gut verschiedene LLMs wie *Gemma 2*, *Llama 3.1* und *ClimateBERT-NetZero* neue freiwillige Unternehmensverpflichtungen zur Dekarbonisierung in US-amerikanischen Nachrichtenartikeln hervorheben können. Untersucht wird, inwieweit diese Modelle unterschiedlicher Komplexität im vorgegebenen semantischen Kontext bestehen können. Zudem soll hierdurch das derzeit ökonomische Potenzial dieser Modelle für das Ableiten mögliche Unternehmensstrategien und Marktanalysen dargestellt werden. Hiermit soll ermöglicht werden, systematische Muster, Ambitionen und die Glaubwürdigkeit solcher Unternehmensversprechen zu analysieren, welches wiederum zu einer besseren Transparenz und Nachvollziehbarkeit beiträgt.

Limitationen ergeben sich, aus der ausgewählten Menge bereits abgeschlossener und trainierter Open-Source Modelle und der genutzten Computerarchitektur. Als Hauptlimitation steht der zur Verfügung stehende Arbeitsspeicher (im Englischen: Random Access Memory, kurz RAM), welcher die Auswahl der Modelle beeinflusst. Diese Arbeitsspeicherbeschränkung von zum Beispiel 32 GB ermöglicht den Einsatz von 8B- und 9B-Modellen wie *Gemma 2*, *Llama 3.1* und *ClimateBERT-NetZero*. Insbesondere soll durch den Vergleich mit dem in der Gesamtheit größeren Modell *GPT-4* die Leistungsfähigkeit der anderen Modelle aufzeigt wer-

den. Als weitere Ebene, soll ein Überblick möglicher Anwendungsbereiche aufgedeckt werden, indem ein auf die kontextspezifische Aufgabe vortrainiertes Modell (*ClimateBERT-NetZero*), den andere allgemeineren vortrainierten Modellen (*Gemma 2* und *Llama 3.1* und *GPT-4*) gegenübergestellt werden.

Eine weitere Herausforderung und zentrales Element der Ergebnisanalyse stellt die Bewertung der semantischen Verständnis- und Ergebnissgüte der LLM-Klassifikationsergebnisse dar. Die Herausforderung besteht darin, die Artikelklassifikationen des Modells zu Bewerten ohne eine objektiv richtige Klassifikation für jeden Artikel zu kennen. Daher werden exemplarisch mithilfe von Experten (im Englischen: Human Coder) Kontrollklassifikationen vorgenommen, um die Leistung automatisierter LLM-Klassifikationen zu validieren.

Final stellt aber auch der Faktor Zeit eine entscheidende limitierende Rolle in der Auswahl und Anpassung der LLMs. Es erfolgt eine bewusste Einschränkung auf derzeit bekannte und verfügbare Modelle, die hinsichtlich ihres kontextspezifischen semantischen Verständnisses und der daraus resultierenden Modellgüte überprüft werden.

1.2 Methodik

Als Datenbasis für die angewandte Sentimentanalyse mittels LLMs werden 1.000 US-amerikanische Nachrichtenartikel, in der Zeitspanne von 2005 bis 2023, klassifiziert, die nach freiwilligen Unternehmensankündigungen einer *neuen*, *klaren*, *umsetzbaren Verpflichtung*, künftige direkte Treibhausgasemissionen (THG) deutlich zu reduzieren, durchsucht werden. Hierzu werden ebenfalls bestehende Reduktionsprojekte mit inbegriffen, sofern neue Informationen bzw. stringentere Ziele verkündet wurden (vgl. Bauer u. a. (Nov 2024)). Diese 1.000 bereitgestellten Artikel entsprechen einem Ausschnitt einer Datensammlung von insgesamt 44.605 Nachrichtenartikeln eines aktuell laufendem Forschungsprojekts mit dem Titel *Corporate Green Pledges* von Bauer u. a. (Nov 2024), das momentan in Kooperation mit Herrn Prof. Dr. Ole Wilms von der Universität Hamburg entsteht. Die Auswahl an Online-Wirtschaftsnachrichten im Datensatz bietet eine umfassende und repräsentative Basis für die Analyse. Hierdurch sollen Einblicke in Unternehmensstrategien zur Dekarbonisierung gewonnen werden und die Leistungsfähigkeit der verwendeten LLMs überprüft werden. Die Klassifizierungsleistungen der LLMs *Gemma 2*, *Lamla 3.1* und *ClimateBERT Netz-Zero* werden mit denen von *GPT-4* verglichen. Bei den genannten LLMs handelt es sich um „unimodale“ LLMs, die ausschließlich Text verarbeiten und generieren. Ebenfalls existiert auch eine multimodale Version des *GPT-4* Modells, bei der das Modell um Bildverarbeitung erweitert ist. Diese Erweiterung ist jedoch für diese Arbeit nicht notwendig, da ausschließlich Textdaten klassifiziert werden.

Durch die Gegenüberstellung der LLMs sollen Erkenntnisse darüber gewonnen werden, inwieweit LLMs mit verschiedener Komplexität und Spezifizierung konkurrieren können. Der Fokus liegt dabei sowohl auf die Genauigkeit der Sentimentanalyse als auch auf deren Effizienz bei der Verarbeitung der bereitgestellten Nachrichtenartikel. Hierzu reichen die besagten unimodalen LLMs, da sie lediglich mit Texteingaben und Textausgaben arbeiten. Der hier relevante Anwendungsbereich (die Klassifizierung von Nachrichtenartikel), spiegelt wiederum nur einen Subbereich des gesamten Spektrums von LLMs ab. Dieser relevante Ausschnitt ist in der folgenden Abbildung 1.1 dargestellt und in roter Farbe hervorgehoben. Die Komplettübersicht von -Kategorisierungen, kann der Abbildung A.1 im Anhang entnommen werden.

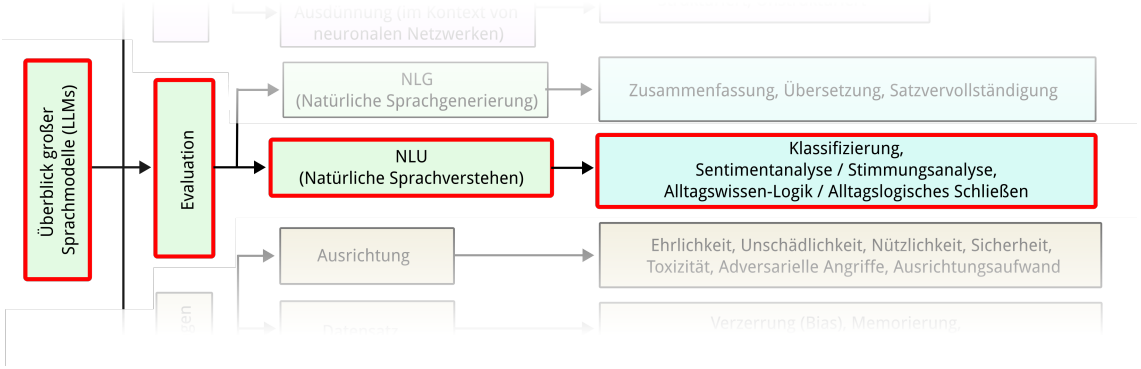

Abbildung 1.1: Auszug des Kernbereichs der LLM Gesamtübersicht.
Eigene Darstellung, in Anlehnung an: Naveed u. a. (2023); S.3.

Die *Open-Source* Modelle *Gemma 2*[1], *Llama 3.1*[2] und *ClimateBERT-NetZero*[3] können für die Klassifizierungsprozesse direkt über das Internet bezogen werden. Bei dem *GPT-4*-Modell hingegen, kann lediglich auf die bereits durch das Forschungsprojekt und Working Paper *Corporate Green Pledges* von Bauer u. a. (Nov 2024) gewonnen Ergebnisse der Klassifizierungen Bezug genommen werden, da dies kein *Open-Source* Modell ist.

[1]Ollama - Gemma 2 Modell https://ollama.com/library/gemma2 In: Ollama, abgerufen am 19. November 2024 um 13:34 Uhr (englisch).
[2]Ollama - Llama 3.1 Modell https://ollama.com/library/llama3.1 In: Ollama, abgerufen am 20. November 2024 um 09:12 Uhr (englisch).
[3]Hugging Face - ClimateBERT-NetZero Modell https://huggingface.co/climatebert/netzero-reduction In: Hugging Face, abgerufen am 20. November 2024 um 08:39 Uhr (englisch).

2 Theoretische Grundlagen

Dieses Kapitel legt die Grundlagen für das Verständnis und die darauffolgende Analyse von LLMs bei der Untersuchung und Klassifizierung freiwilliger Dekarbonisierungsverpflichtungen. Es behandelt die grundlegenden Kernkonzepte des maschinellen Lernens, einschließlich einer Einführung der Nutzung von LLMs in der Ökonomie und ihrer Lernstile. Außerdem werden Verfahren der Textklassifizierung, des Text Minings und der Computerlinguistik erklärt, die für die Verarbeitung umfangreicher Textdatensätze entscheidend sind um im anschließenden Abschnitt 3 mithilfe von Beispielen weiter vertieft werden.

2.1 Maschinelles Lernen

Maschinelles Lernen, das einem Bereich der Künstliche Intelligenz (KI) zugeordnet werden kann, ist ein Teilgebiet der Informatik. Es kann in bestehenden Datensätzen vordefinierte Muster identifizieren und die dabei gewonnenen Informationen auf weitere Daten mit ähnlicher Struktur Anwenden. Hierdurch können neue oder effizientere Lösungen entwickeln werden. Ein KI-Algorithmus zur Entwicklung einer Spracherkennung kann nur dann entwickelt werden, wenn zuvor geeignete Lernalgorithmen verwendet wurden. (vgl. Russell und Norvig (2012)).

2.1.1 Einsatz von LLMs in der Ökonomie und Finanzwirtschaft

Mit der immer schneller voranschreitenden Entwicklung von LLMs hat sich das Potenzial dieser Technologien auch in den Naturwissenschaften sowie in der Ökonomie und Finanzwirtschaft erheblich ausgeweitet.

Der Entwicklungsfortschritt im Bereich der LLMs lässt sich anhand zweier aufschlussreicher Abbildungen aus der Studie von Naveed u. a. (2023) gut nachvollziehen. Diese Grafiken bieten einen detaillierten Einblick in die zeitliche Entwicklung und die Charakteristika der LLM-Forschung. Abbildung A.2 illustriert den zeitlichen Verlauf von Publikationen und Anwendungen im Kontext von LLMs. Hierbei wird der Fokus auf drei Schlüsselbegriffe gelegt: *LLM*, *LLM+Fine-Tuning* und *LLM+Alignment*. Die Abbildung ermöglicht es, den rasanten ansteigenden Trend wissenschaftlicher Veröffentlichungen in diesem Forschungsfeld über die Jahre hinweg nachzuvollziehen. Ein weiterer bedeutsamer Aspekt wird in Abbildung A.3 aufgezeigt. Diese Abbildung zeigt die Entwicklung des Verhältnisses zwischen Open-Source- und Closed-Source-LLM-Modellen im Zeitverlauf. Die Visualisierung unterscheidet dabei zwischen vortrainierten (blau dargestellt) und instruiert-trainierten Modellen (orange dargestellt). Die obere Hälfte der Abbildung repräsentiert Open-Source-Modelle, während die untere Hälfte Closed-

Source-Modelle abbildet. Besonders hervorzuheben ist der erkennbare Trend hin zu instruiert-trainierten und Open-Source-Modellen, was die dynamische Entwicklung in der Forschungslandschaft der natürlichen Sprachverarbeitung unterstreicht. Heutzutage sind diese LLMs schon in der Lage, Anweisungen in fachspezifischer und umgangssprachlicher Sprache präzise und effizient zu verstehen und umzusetzen. Dadurch bietet sie eine breite Palette von Anwendungsmöglichkeiten in den unterschiedlichsten ökonomischen Szenarien und Aufgabenbereichen. Insbesondere lassen sich die Einsatzgebiete von LLMs in der Finanzwirtschaft in vier zentrale Kategorien einteilen, die maßgeblich die aktuellen Anwendungen prägen (vgl. H. Zhao u. a. (2024)):

Finanzengineering

Im Finanzengineering kommen LLMs zum Einsatz, um große Datenmengen auszuwerten und komplexe Finanzinstrumente sowie -Strategien zu bearbeiten. Hierzu zählen die Modellierung strukturierter Produkte, die Bewertung von Derivaten oder die Optimierung von Portfolios. Sie erlauben außerdem die Automatisierung komplexer Berechnungen, für die man herkömmlicherweise große Rechenressourcen braucht.

Finanzprognosen

LLMs tragen im Bereich der Finanzprognosen unter anderem dazu bei, Markttrends vorherzusagen und gut begründete Entscheidungen zu Investitionen abzuleiten. Ebenfalls werden dank der Modelle, tiefergehende Analysen des Konsumverhaltens, der Marktdynamiken oder verschiedener Segmentierungen ermöglicht. Außerdem bieten LLMs die Möglichkeit, auf Marktbewegungen schneller reagieren zu können, indem sie Finanznachrichten und zum Beispiel Social-Media-Inhalte in Echtzeit auswerten.

Finanzrisikomanagement

LLMs helfen durch die Analyse umfangreicher Daten dabei, potenzielle Risiken frühzeitig zu identifizieren und zu bewerten. Dies umfasst beispielsweise die Bewertung des Kreditrisikos in Banken, die Untersuchung von Marktrisiken und die Überprüfung von Betrugsmustern. LLMs sind aufgrund ihrer besonderen Eignung für die Echtzeitentdeckung von Anomalien im Risikomanagement besonders wertvoll.

Echtzeit-Fragenbeantwortung im Finanzbereich

LLMs ermöglichen die präzise und schnelle Beantwortung von Finanzfragen in Echtzeit, was insbesondere für den Kundenservice und die datenbasierte Entscheidungsfindung von großer Bedeutung ist. Ein Beispiel ist die dynamische

Preisoptimierung im E-Commerce: LLMs analysieren hier in Echtzeit Marktdaten, Kundenverhalten und Wettbewerbsinformationen, um Preise automatisch anzupassen und optimale Verkaufsergebnisse zu erzielen. Solche Anwendungen steigern nicht nur die Effizienz, sondern schaffen auch Wettbewerbsvorteile durch eine verbesserte Reaktionsfähigkeit.

Die Studie von H. Zhao u. a. (2024) demonstriert durch die Integration von LLMs im Finanzsektor, das ein relevantes Potenzial durch ihrer Einbindung aufgezeigt werden kann. Sie heben hervor, dass eine signifikante Steigerung der Effizienz, fundiertere Entscheidungsfindungen und eine verbesserte Kundenzufriedenheit geschaffen werden könnte. *GPT-4*-Tests zeigten, dass das Modell Anweisungen in Umgangssprache effektiv umsetzen konnte und entsprechend vielseitig in den verschiedensten finanziellen Sektoren einsetzbar ist. Gleichzeitig weist die Studie aber auf Herausforderungen hin, insbesondere die Komplexität der Daten und die hohen Anforderungen an Genauigkeit und Zuverlässigkeit. Es wird empfohlen, spezialisierte Trainingsdaten mit Expertenwissen zu kombinieren, um die Leistungsfähigkeit der Modelle zu maximieren. Diese Herangehensweise wird ebenfalls ein fundamentaler Aspekt bei der Auswertung der hier entstehenden Ergebnisse der angewandten Experimente sein. Resultierende Erkenntnisse unterstreichen die Relevanz von LLMs und lassen sich auf den Einsatz in klimabezogenen Analysen übertragen, wo ebenfalls komplexe Daten und präzise Ergebnisse entscheidend sind.

2.1.2 Lernstile des maschinellen Lernens

Maschinelles Lernen beinhaltet verschiedene Lernstile, wobei überwachtes und unüberwachtes Lernen die Hauptmethoden sind. Diese Lernansätze sind die Basis für moderne Modelle, einschließlich LLMs, die immer häufiger in Anwendungen wie Textklassifikationen verwendet werden.

Überwachtes Lernen

Beim überwachten Lernen wird ein Modell mit klassifizierten Daten trainiert, wobei Eingaben mit Zielwerten (Klassen, im Englischen: *Labels*) verknüpft sind. Es eignet sich ideal für Aufgaben wie Sentimentanalysen, da klare Zielvariablen vorgegeben werden (vgl. Döbel u. a. (2018)). Überwachtes Lernen zielt darauf ab, die Beziehung zwischen Eingabemerkmalen und den zugehörigen Klassen zu modellieren, um ein Modell zu entwickeln, das Klassen für unbekannte Daten vorhersagen kann. Der Prozess umfasst die Merkmalsextraktion aus Trainingsdaten, das Modelltraining zur Anpassung an die zugrunde liegenden Zusammenhänge und die Modellbewertung anhand separater Validierungsdaten. Je nach Genauigkeit wird der Prozess iterativ optimiert, bevor das trainierte Modell auf neue

Daten angewendet wird, um Klassen vorherzusagen (vgl. Raschka und Mirjalili (2019)).

Unüberwachtes Lernen

Im Gegensatz dazu identifiziert das unüberwachte Lernen Muster und Strukturen in unklassifizierten Daten, etwa durch die Zusammenfassung „ähnlicher" Klassifizierungen in Gruppen (im Englischen: Clustering) oder mittels Dimensionsreduktion (vgl. Döbel u. a. (2018)). Dieser Lernstil findet oft seinen Einsatz bei sehr großen, unstrukturierte Datenmengen, bei denen es im Vorfeld nicht bekannt ist, wie gut sie beschrieben oder nach welchen Kriterien sie aufgeteilt werden sollen (vgl. Döbel u. a. (2018); S.26).

Gesamt betrachtet, vereinen LLMs beide Herangehensweisen: In der Vor-Trainingsphase wird häufig unüberwachtes Lernen verwendet, um Textdaten im großen Umfang zu verarbeiten und ein kontextspezifisches Sprachverständnis aufzubauen. Anschließend spielt das überwachte Lernen eine entscheidende Rolle dabei, spezielle Aufgaben wie Textklassifikationen zu realisieren.

2.2 Text Mining / Computerlinguistik

Text Mining dient der systematischen Analyse unstrukturierter Textdaten mit dem Ziel, informative Muster und Erkenntnisse zu gewinnen. Hierbei kommen Methoden der Computerlinguistik und des maschinellen Lernens zum Einsatz, um Texte zu klassifizieren, Gruppen zuzuordnen (*clustern*) oder Schlüsselthemen herauszufiltern. Ein gutes Beispiel für dieses Anwendungsfeld sind Nachrichtenartikel, da sie große Mengen unstrukturierter Daten enthalten, die durch Text Mining effizient analysiert und auf relevante Inhalte reduziert werden können. Obwohl die systematische Analyse unstrukturierter Textdaten durch Text Mining möglich ist, werden leistungsfähige Modelle benötigt, um komplexe kontextuelle Abhängigkeiten zu identifizieren. Die traditionellen Methoden des Text Mining und der Computerlinguistik haben die Basis für die Automatisierung der Verarbeitung natürlicher Sprache geschaffen. Transformer-Modelle hingegen stellen einen bedeutenden Wendepunkt in diesem Forschungsbereich dar. Die Architekturen im anschließenden Abschnitt hat die Leistungsfähigkeit von Systemen revolutioniert, welche die Verarbeitung natürlichsprachlicher Informationen im fokus haben und neue Möglichkeiten bieten noch komplexere sprachbasierte Aufgaben behandeln zu können. Diese sogenannte Transformer-Architektur basiert auf den Erkenntnissen früherer Ansätze, geht jedoch weit darüber hinaus, indem sie effizientere Methoden zur Erfassung von Kontextinformationen und sprachlichen Nuancen einführt.

2.3 Transformermodelle

Transformer-Modelle repräsentieren den Stand der Technik in der Verarbeitung natürlicher Sprache und zeichnen sich durch ihre Fähigkeit aus, komplexe Muster in sequenziellen Daten effizient zu lernen. Durch den Einsatz von tiefen neuronalen Netzen zur Erfassung kontextueller Abhängigkeiten in Texten wird die Modellierung nicht-linearer Beziehungen in großen Datensätzen ermöglicht (vgl. Raschka und Mirjalili (2019)). Die folgenden grundlegenden Mechanismen tiefer neuronaler Netze sind hierbei entscheidend:

Selbstaufmerksamkeit (Self-Attention)

Ein wesentlicher Mechanismus der Transformermodelle ist die Selbstaufmerksamkeit. Dieser Mechanismus ermöglicht es, die Beziehungen zwischen verschiedenen Wörtern oder Wortteilen innerhalb eines Textes zu erkennen, unabhängig von deren Entfernung im Satz. Dies ermöglicht die Gewichtung relevanter Kontextinformationen unabhängig von der Eingabelänge. Dieser Ansatz erlaubt es dem Modell, sowohl lokale als auch globale Bedeutungsstrukturen im Text zu erfassen und verbessert dadurch die Qualität der Textverarbeitung und -Generierung erheblich. (vgl. Raschka und Mirjalili (2019); S. 613 – 618)

Sequenzielle Datenverarbeitung

Transformermodelle und rekurrente neuronale Netze (RNNs) unterscheiden sich grundlegend in ihrer Architektur und Funktionsweise zur Verarbeitung von Eingabesequenzen. Während RNNs Wort- oder Satzteile sequenziell verarbeiten und dabei Informationen durch Schleifenstruktur über Zeit hinweg weitergeben, ermöglichen Transformermodelle durch den Mechanismus der Selbstaufmerksamkeit die parallele Verarbeitung aller Wort- oder Satzteile eines Textes, wodurch sie wesentlich effizienter sind. Dies erlaubt Transformermodellen, sowohl kurze als auch weit entfernte Abhängigkeiten innerhalb eines Textes besser zu erfassen, während RNNs bei langen Sequenzen häufig Probleme mit dem sogenannten *Vanishing-Gradienten*-Problem haben. Das Vanishing-Gradienten-Problem beschreibt ein Phänomen in tiefen neuronalen Netzen, insbesondere bei rekurrenten Architekturen wie RNNs, bei dem die Werte, die die Anpassung der Modellparameter steuern, während der Rückwärtsberechnung von der Ausgabeschicht zur Eingabeschicht stark abnehmen. Dadurch werden die Parameter in den frühen Schichten des Netzes kaum noch verändert, was die Fähigkeit des Modells, aus den Daten zu lernen, deutlich einschränkt. (vgl. Raschka und Mirjalili (2019))

Transformermodelle verbinden durch diese Mechanismen hohe Modellierungsfähigkeit mit Rechenleistung und haben die Verarbeitung natürlicher Sprache nachhaltig vorangetrieben. Ihr Können im effektiven Modellieren von komple-

xen und sequenziellen Daten hebt sie entsprechend hervor und für natürliche Sprachverstehen oder vergleichbare Aufgaben sind sie mittlerweile zu einem Standard geworden. Dagegen bieten traditionelle Klassifikationsalgorithmen wie die k-Nächste-Nachbarn-Klassifikation (kNN), Random Forest (RF) und Support-Vektor-Maschine (SVM) robuste, einfache und ressourcenschonende Lösungen für kleinere und tabellarische Datensätze. Ihre Entscheidung ist abhängig vom spezifischen Anwendungsfall. Die Transformer-Modelle, die in dieser Arbeit verwendet werden, stehen im starken Gegensatz zu klassischen Klassifikationsalgorithmen. Dies resultiert aus grundlegenden Differenzen in Architektur, Funktionsweise und Anwendungsgebieten.

3 Modellentwicklung und -Implementierung

Dieses Kapitel beschreibt die zentralen Schritte der Modellentwicklung und -Implementierung der anschließend angewandten Experimente und Analysen freiwilliger Dekarbonisierungsverpflichtungen in Nachrichtenartikeln. Zunächst werden die Datenbasis sowie die Datenanalyse und -Reduktion von Datendimensionen dargestellt, um eine verlässliche Grundlage zu schaffen. In der anschließenden Phase der Datenvorverarbeitung liegt der Schwerpunkt auf der Textvorverarbeitung, einschließlich Tokenisierung und Bereinigung. Um die Daten bestmöglich auf eine Modellauswertung vorzubereiten, werden im nächsten Schritt Transformationstechniken, einschließlich Text-Transformationen und Dimensionsreduktion, erklärt. Dieses strukturierte Vorgehen bildet die Grundlage für eine anschließende Analyse und Interpretation von Modellergebnisse. Hierbei liegt der Schwerpunkt auf Klassifikationsmethoden und der Bewertung anhand von Metriken binärer Klassifikationen, um die Effektivität der Modelle in diesem Abschnitt der Analyse besser einordnen zu können.

3.1 Datengrundlage

Basis dieser Arbeit bildet ein neu angelegter Datensatz. Dieser erfasst zeitlich protokollierte Reduzierungen von Treibhausgasemissionen sowie *grüne Versprechen* öffentlicher US-Unternehmen. Der Datensatz mit Ankündigungen freiwilliger Unternehmensverpflichtungen wurde im Rahmen der Forschungsarbeit mit dem Working Paper-Titel *Corporate Green Pledges* von Bauer u. a. (Nov 2024) erstellt. Er gründet sich auf eine systematische Analyse eines umfangreichen Korpus von Nachrichtenartikeln, der durch das Extrahieren von Inhalten und Daten verschiedener Webseiten gewonnen wurde. Der Datensatz deckt den Zeitraum von 2005 bis 2023 ab und beinhaltet eine Vielzahl von wirtschaftsbezogenen Unternehmensnachrichten, darunter Pressemitteilungen, Gewinnankündigungen, Unternehmensberichte sowie andere Formen der Unternehmenskommunikation (vgl. Bauer u. a. (Nov 2024)).

3.2 Datenanalyse

Die Identifikation der relevanten Ankündigungen erfolgte durch die Kombination manueller Codierung und der Unterstützung eines großen Sprachmodells. Dies ermöglichte die Extraktion zeitlich markierter Daten und die Kategorisierung der Dekarbonisierungszusagen, die als Grundlage für die nachfolgende Analyse dienen. Dieser Ansatz zeigt die Wirksamkeit moderner Methoden zur

systematischen Verarbeitung großer Mengen unstrukturierter Daten und deren Transformation in strukturierte, analysierbare Datensätze.

3.3 Data Mining

Data Mining ist ein Verfahren zur Entdeckung verborgener Muster und Zusammenhänge in strukturierten Daten, wie Datenbanken oder tabellarischen Formaten. Es bildet die Grundlage für viele Analyseprozesse, einschließlich der Datenvorbereitung und Mustererkennung, die auch im Text Mining Anwendung finden. Nachrichtenartikel, obwohl unstrukturiert, erfordern ähnliche Schritte wie Data Mining, z. B. Vorverarbeitung und Merkmalsextraktion, bevor sie effektiv analysiert werden können. So verbindet Text Mining die Prinzipien von Data Mining mit spezifischen Ansätzen zur Analyse von Textdaten.

3.4 Datenvorverarbeitung

Die Datenvorverarbeitung ist ein zentraler Schritt, um unstrukturierte Textdaten für die Analyse vorzubereiten. Zunächst werden durch *Tokenisierung* Texte in kleinere Einheiten wie Wörter zerlegt, bevor mithilfe der *Lemmatisierung* Wörter auf ihre Grundform reduziert werden, um semantische Konsistenz zu schaffen. Anschließend eliminiert die *Stopword*-Entfernung wenig informative Wörter, wodurch der Fokus auf relevante Inhalte gelegt wird. Abschließend werden *N*-Grams genutzt, um Sequenzen mehrerer Wörter zu erfassen und kontextuelle Muster im Text zu identifizieren. Diese Schritte schaffen eine Grundlage für die effiziente Nutzung moderner Klassifikationsmodelle.

Textvorverarbeitung

Der folgende Abschnitt zeigt anhand eines exemplarischen Beispiels auf, wie LLMs in Kombination verschiedener möglicher Textanalyseansätz dazu befähigt sind, eine umfassendere Sentimentanalyse vorzunehmen. Zunächst wird der Tokenisierungsprozess von Text zu numerischen Werten aufgezeigt und anschließend weitere Techniken wie ein LLM sowohl einzelne Wörter als auch Wortsequenzen berücksichtigen kann.

Exemplarischer GPT-4o tokenisierungsprozess

Abbildung 3.1 und die anschließende Auflistung resultierender Zahlenwerte, zeigen einen exemplarischen numerischen Tokenisierungsprozess auf, anhang eines Original-Artikelauszugs (Artikel 982 von 1.000) des Datensatzes von Bauer u. a. (Nov 2024). In Abbildung 3.1 lassen sich mithilfe Hilfe der jeweiligen Farbzuwei-

sungen die Überführungen von Textelement zu den anschließenden numerischen Werten gut abbilden.

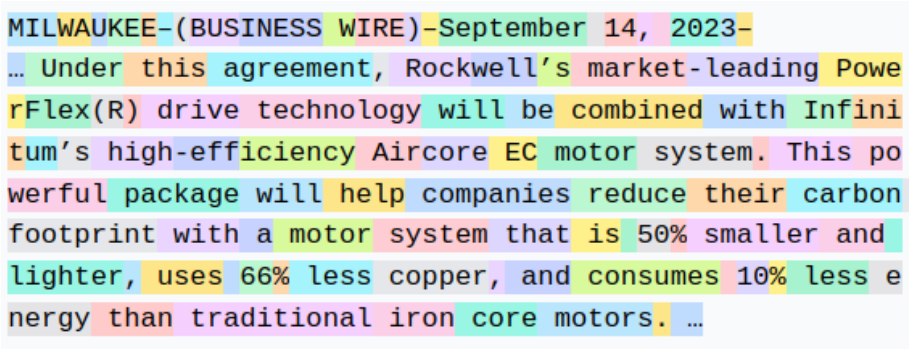

Abbildung 3.1: Exemplarischer GPT-4o Tokenizer Prozess anhand eines originalen Artikelauszugs (Artikel 982 von 1.000) des Datensatzes von Bauer u. a. (Nov 2024). Eigene Darstellung, in Anlehnung an: https://tiktokenizer.vercel.app/ (Abgerufen am 31.12.2024 um 16:58 Uhr))

```
[181132, 26913, 52, 6003, 36, 1585, 7, 71740, 43601, 486, 12934, 8,
1585, 30538, 220, 1265, 11, 220, 1323, 18, 170523, 1131, 13911, 495,
15072, 11, 12251, 13072, 802, 2910, 92557, 10079, 37730, 4092, 8, 7515,
6340, 738, 413, 15890, 483, 17857, 3287, 394, 802, 1932, 118351, 12895,
6625, 6978, 34789, 9771, 2420, 13, 1328, 11629, 9726, 738, 1652, 6005,
10389, 1043, 15883, 62896, 483, 261, 9771, 2420, 484, 382, 220, 1434,
4, 13679, 326, 44854, 11, 8844, 220, 3618, 4, 3760, 34855, 11, 326,
109297, 220, 702, 4, 3760, 5954, 1572, 10634, 17069, 10089, 54380,
13, 3762]
```

Des Weiteren nutzen LLMs aber auch Textvorverarbeitungstechniken wie beispielsweise Lemmatisierung, Tokenisierung und Stopwords Entfernung, welches zu einer Dimensionsreduzierung führt. Diese Schritte werden ebenfalls exemplarischen anhand des zuvor aufgeführten Artikelauszugs und in Anlehnung von Işik und Dağ (2020) Überführungen aufgezeigt.

Lemmatisierung

Ein wesentlicher Schritt in der Textvorverarbeitung ist die Lemmatisierung. Sie soll Wörter auf ihre Grundform (Lemma) zurückführen. Lemmatisierung berücksichtigt im Gegensatz zur bloßen Stammerkennung (Stemming) grammatikalische und kontextuelle Informationen, um die semantische Bedeutung der Wörter zu bewahren. Die Wörter „running“ und „ran“ werden zum Beispiel auf das Lemma „run“ zurückgeführt. Hierdurch wird eine wirksamere Normalisierung möglich und die Qualität der nachfolgenden Analysen in Modellen wie den Transformer-Architekturen verbessert. Die Lemmatisierung gewährleistet eine konsistente Darstellung verwandter Wörter, was besonders bei der Tokenisierung für genauere Ergebnisse entscheidend ist.

Tokenisierung

Tokenisierung ist ein methodischer Schritt, bei dem ein zusammenhängender Text in kleinere Einheiten (*Tokens*) zerlegt wird, um ihn für maschinelle Lernverfahren verarbeitbar zu machen. Diese Einheiten können Wörter, Subwörter oder sogar einzelne Zeichen sein, abhängig von der Granularität, die das Modell erfordert. In Transformermodellen ist die Tokenisierung entscheidend, da sie als erste Verarbeitungsebene fungiert und sicherstellt, dass der Text in eine für das Modell verständliche numerische Form umgewandelt wird. Der Original-Artikelauszug (Artikel 982 von 1.000) des Datensatzes von Bauer u. a. (Nov 2024) soll weiterhin exemplarisch dienen den Tokenisierungsprozess besser verstehen zu können:

Original-Artikelauszug (Artikel 982 von 1.000):

> MILWAUKEE--(BUSINESS WIRE)--September 14, 2023--
> … Under this agreement, Rockwell's market-leading PowerFlex(R) drive technology will be combined with Infinitum's high-efficiency Aircore EC motor system. This powerful package will help companies reduce their carbon footprint with a motor system that is 50% smaller and lighter, uses 66% less copper, and consumes 10% less energy than traditional iron core motors. …

wird zu der tokenisierten Textform:

```
'milwaukee', '--', '(', 'business', 'wire', ')', '--', 'september',
'14', ',', '2023', '--', 'under', 'this', 'agreement', ',', 'rockwell',
„'s“, 'market', '-', 'leading', 'powerflex', '(', 'r', ')', 'drive',
'technology', 'will', 'be', 'combined', 'with', 'infinitum', „'s“,
'high', '-', 'efficiency', 'aircore', 'ec', 'motor', 'system', '.',
'this', 'powerful', 'package', 'will', 'help', 'companies', 'reduce',
'their', 'carbon', 'footprint', 'with', 'a', 'motor', 'system', 'that',
'is', '50', '%', 'smaller', 'and', 'lighter', ',', 'uses', '66', '%',
'less', 'copper', ',', 'and', 'consumes', '10', '%', 'less', 'energy',
'than', 'traditional', 'iron', 'core', 'motors', '.'
```

Transformer-Modelle wie *GPT-4* verwenden Tokenisierung als Vorverarbeitungsschritt, um die Texteingabe in Vektoren umzuwandeln. Diese numerischen Repräsentationen bilden die Grundlage für Mechanismen wie Selbstaufmerksamkeit, die semantische und syntaktische Beziehungen innerhalb des Textes analysieren. Der hier gezeigte Schritt ist somit essenziell, um die strukturelle und kontextuelle Bedeutung des Textes modellieren zu können.

Stopwords Entfernung

Die Stopword-Entfernung eliminiert häufig vorkommende Wörter wie „und“, „oder“ und „ist“, die in der Regel wenig semantischen Mehrwert für die Analyse bieten. Um die Anforderungen an Verarbeitung und Speicherung zu verringern, werden diese Wörter unter Verwendung von Stopword-Listen oder spezifischen Filter für die Domäne entfernt. Dies trägt dazu bei, dass das Modell aussagekräftige Terme in den Blick nimmt und die Effektivität der nachfolgenden Analy-

seschritte steigert. Die Entfernung der Stopwords ist bei Transformer-Modellen nicht zwingend erforderlich, da sie durch ihre Kontextsensitivität oft in der Lage sind, auch mit Stopwords umzugehen. Dennoch kann sie dazu beitragen, die Effizienz bei der Verarbeitung großer Datenmengen zu steigern.

Tokenisierter Artikel ohne Stopwörter:

```
['milwaukee', '--', '(', 'business', 'wire', ')', '--', 'september',
'14', ',', '2023', '--', 'agreement', ',', 'rockwell', „'s“, 'market',
'-', 'leading', 'powerflex', '(', 'r', ')', 'drive', 'technology',
'combined', 'infinitum', „'s“, 'high', '-', 'efficiency', 'aircore',
'ec', 'motor', '.', 'powerful', 'package', 'help', 'companies', 'reduce',
'carbon', 'footprint', 'motor', '50', '%', 'smaller', 'lighter', ',',
'uses', '66', '%', 'copper', ',', 'consumes', '10', '%', 'energy',
'traditional', 'iron', 'core', 'motors', '.']
```

3.5 Embedding

Im Anschluss an der Tokenisierung nutzen Transformermodelle sogenannte *Embedding*-Methoden, bei denen Wörter numerisch in einem mehrdimensionalen Raum repräsentiert werden. Ziel ist es, die semantische Ähnlichkeit von Wörtern durch die Nähe ihrer Vektoren in diesem Raum abzubilden. Die Werte in den Vektoren basieren auf der Interbeziehung zwischen Wörtern: Je stärker ihre semantische Ähnlichkeit, desto höher ihr Wert (maximal 1); je weiter sie voneinander entfernt sind, desto geringer ihr Wert (minimal -1). Die folgende Abbildung 3.2 bietet einen anschaulichen Einblick in diesen Prozess.

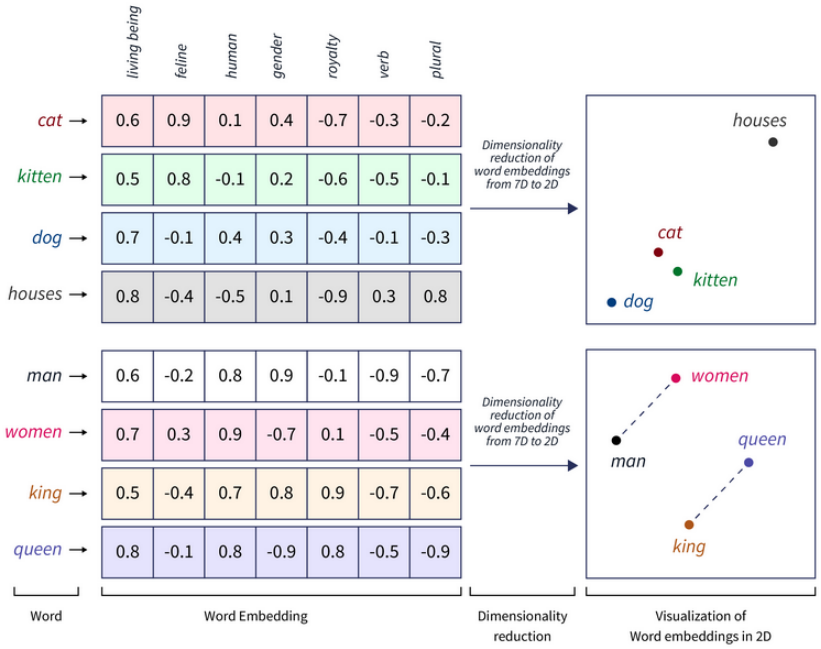

Abbildung 3.2: Exemplarischer Embeddingprozess. Quelle: https://www.scaler.com/topics/tensorflow/tensorflow-word-embeddings/ (Abgerufen am 28.12.2024 um 17:22 Uhr)

3.6 Textklassifikationsverfahren

Klassifikationsprobleme lassen sich in drei Typen unterteilen: *Ein-Klassen-* (*binäre*), *Mehr-Klassen-* und *Multi-Klassen-Klassifikationen*. Diese Typarten sind in der folgenden Abbildung 3.3 schematisch dargestellt.

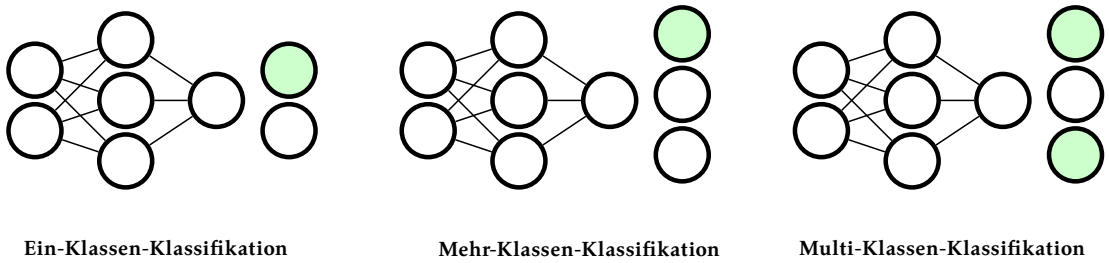

Abbildung 3.3: Übersicht der Klassen-Klassifikationstypen. Eigene Darstellung, in Anlehnung an: https://huggingface.co/blog/sirluk/multilabel-llm (Abgerufen am 05.12.2024 um 09:22 Uhr)

Ein-Klassen-Klassifikation

Eine Ein-Klassen-Klassifikation, oder auch binäre Klassifizierung, beschreibt eine Klassen-Klassifizierung zwischen zwei sich gegenseitig ausschließenden Kategorieoptionen, wie z.B. der Identifikation von einer Zustandsbewertung mit 1 für zutreffend oder 0 für nicht zutreffend. Im Abschnitt 4.2 *Experimente* werden die Modelle *Gemma 2*, *Llama 3.1* und *GPT-4* dieser Klassifizierungsform entsprechen.

Mehr-Klassen-Klassifikation

Eine Mehr-Klassen-Klassifikation ist eine Erweiterung der Ein-Klassen-Klassifikation, bei der eine Kategorie aus mehreren möglichen, oft gegenseitig ausschließenden Kategorieoptionen ausgewählt wird, wie beispielsweise bei der Text- oder Bildklassifikation. Im Fall von Mehr-Klassen-Klassifikationen mit LLMs, wie etwa dem *ClimateBERT-NetZero*-Modell mit den Klassen $L = \{\text{Ohne Bezug,-Reduktion, Net-Zero}\}$, kann es jedoch vorkommen, dass die Kategorien nicht strikt gegenseitig ausschließend sind. Dies liegt daran, dass die semantische Abgrenzung der Klassen vom spezifischen Kontext oder Interpretationsspielraum abhängt. Im Abschnitt 4.2 *Experimente* wird *ClimateBERT-NetZero* als Beispiel für diese Klassifizierungsform vorgestellt.

Multi-Klassen-Klassifikation

Bei einer Multi-Klassen-Klassifikation hingegen, wird die Bedingung der gegenseitigen Ausschließlichkeit ganz aufgehoben, sodass ein Artikel gleichzeitig mehreren Klassen zugeordnet werden kann. Dies sorgt für eine genauere Klassifizierung, falls mehrere Kategorien gleichermaßen zutreffend sind. Beispiel hierfür

kann die Kennzeichnung eines Artikels mit sowohl „`Reduktion`“ als auch „`Net-Zero`“ sein.

3.7 Bewertung von Klassifikationen

Die Evaluierungsmaße für Mehr-Klassen-Klassifikation unterscheiden sich von denen, die für die Ein-Klassen-Klassifikation (binäre Klassifikationsergebnisse) verwendet werden. Evaluierungsmaße fallen in zwei Kategorien: *labelbasierte* und *beispielbasierte*. Labelbasierte Messungen sind eine erweiterte Form der Evaluierungsmessungen, die im Bereich der Ein-Klassen-Klassifikation verwendet werden. Beispielbasierte Messungen sind speziell für den Mehrklassenbereich aufgestellt worden (vgl. Maimon und Rokach (2010)).

In den folgenden Evaluierungsformeln für Ein-Klassen-Klassifikation entspricht jeweils das x dem vom KNN-Klassifikator vorhergesagte Klasse und das y dem der richtigen Klasse. Des Weiteren entspricht N der Gesamtmenge an Nachrichtenartikeln (vgl. Asim, Rehman und Shoaib (2017); S. 374).

Parallel hierzu bietet es sich an, ein Teil der folgenden Gleichungen ebenfalls in Hinsicht auf Richtig- und Fehlklassifikationen zu beleuchten. Hiermit lässt sich später im Rückschluss mithilfe der quantitativen Metriken eine Konfusionsmatrix zu jedem (Binären) verwendeten LLM bilden, um Rückschluss auf die jeweilige semantische Genauigkeit im Verhältnis zu Bewerten. Hierbei können die folgenden vier Fälle auftreten (vgl. Asim, Rehman und Shoaib (2017); S. 375):

- Falsch negativ (f_n): Fälle, in denen das Modell falsch „Nein“ klassifiziert.
- Falsch positiv (f_p): Fälle, in denen das Modell falsch „Ja“ klassifiziert.
- Richtig negativ (r_n): Fälle, in denen das Modell korrekt „Nein“ klassifiziert.
- Richtig positiv (r_p): Fälle, in denen das Modell korrekt „Ja“ klassifiziert.

Korrektklassifikationsrate (Treffergenauigkeit)

Die Korrektklassifikationsrate (englisch: Accuracy) kann als eine bedingte Wahrscheinlichkeit interpretiert werden und gibt den Anteil an korrekt vorhergesagten Klassifikationen im Verhältnis zur Gesamtzahl der Vorhersagen an. Sie wird als Verhältnis der richtigen Vorhersagen zu den gesamten Vorhersagen berechnet und eignet sich gut für ausgewogene Datensätze.

$$P(\text{Modell richtig positiv}) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i \bigcap y_i}{x_i \bigcup y_i}\right| \quad \text{bzw.} \quad \frac{r_p + r_n}{r_p + r_n + f_p + f_n}$$

Positiver Vorhersagewert (Genauigkeit)

Der positive Vorhersagewert (englisch: Precision) gibt die bedingte Wahrscheinlichkeit an, wie viele der vom Modell als positiv klassifizierten Beispiele tatsächlich positiv sind. Sie wird als Verhältnis der richtig positiven Vorhersagen zu der Gesamtzahl der als positiv klassifizierten Beispiele berechnet.

$$P(\text{Artikel positiv} \mid \text{Modell positiv}) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{x_i \cap y_i}{x_i} \right| \quad \text{bzw.} \quad \frac{r_p}{r_p + f_p}$$

Richtig-positiv-Rate (Sensitivität)

Die Richtig-positiv-Rate (englisch: Recall) misst den Anteil der tatsächlich positiven Beispiele, die korrekt als positiv identifiziert wurden. Er wird als Verhältnis der richtig positiven Vorhersagen zur Gesamtzahl der tatsächlichen positiven Beispiele berechnet.

$$P(\text{Artikel positiv} \mid \text{Modell richtig positiv}) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{x_i \cap y_i}{y_i} \right| \quad \text{bzw.} \quad \frac{r_p}{r_p + f_n}$$

F-Maß

Das F-Maß (englisch: F_1 score) ist das harmonische Mittel bei dem Precision und Recall gleich gewichtet sind, um eine ausgewogene Metrik zu bilden. Das F-Maß wird verwendet, um das Gleichgewicht zwischen diesen beiden Metriken zu bewerten, welches insbesondere bei ungleichen Klassenzahlen Relevanz gewinnt.

$$\text{F-Maß} = \frac{1}{N} \sum_{i=1}^{N} 2 \cdot \frac{|x_i \cap y_i|}{|x_i| + |y_i|} \quad \text{bzw.} \quad 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Richtig-negativ-Rate (Spezifität)

Die Richtig-negativ-Rate (englisch: Specificity) zeigt, wie gut das Modell irrelevante oder nicht zutreffende semantische Inhalte von den relevanten unterscheidet. Eine hohe Spezifität bedeutet, dass das Modell fehlerhafte positive Klassifikationen minimiert und somit unnötige oder irrelevante Ergebnisse vermeidet.

$$P(\text{Artikel negativ} \mid \text{Modell richtig negativ}) = \frac{r_n}{r_n + f_p}$$

4 Modellanalyse und -Ergebnisse

Die Qualität der LLM-Klassifikationen hängt von der Fähigkeit des Modells ab, relevante Inhalte korrekt zu interpretieren und die Zielvorgaben präzise umzusetzen. Es ist jedoch immer mit einem gewissen Fehlerquotienten zu rechnen, vor allem in Gestalt von Fehlinterpretationen oder Fehlklassifikationen. Ziel sollte es daher sein, die Fehlerquote so gering wie möglich zu halten, um die Verlässlichkeit des Modells zu optimieren. Der Vorteil von vortrainierten Spezialmodellen wie *ClimateBERT-NetZero*, die auf thematisch relevanten Datensätzen wie Klima- und Umwelttexten trainiert wurden, ist, dass sie in ihrem Fachbereich eine tiefere und kontextsensitive Klassifizierung ermöglichen. Dies kann Ergebnisse mit höherer Genauigkeit zur Folge haben, da die Modelle bereits über spezifisches Wissen und Muster aus dem Anwendungsgebiet verfügen. Dagegen weisen größere, allgemein ausgerichtete Modelle wie *GPT-4* eine breitere Anwendbarkeit auf, sind jedoch nicht speziell auf die Feinheiten eines bestimmten Themenbereichs abgestimmt. Dies kann dazu führen, dass Entscheidungen stärker verallgemeinert werden und spezifische Nuancen des Anwendungsbereichs übersehen werden.

Ebenfalls relevant ist die Form der Sprache, welche sich je nach Autor und Zeitepoche stets wandelt und eine weitere kontinuierliche Fehlerquelle für richtiges Klassifizieren darstellen kann. Diese feinen sprachlichen Nuancen, können wiederum spezielles Fachwissen bedeuten, was von einem generell größerem Modell wie *GPT-4* möglicherweise passendere Interpretationen zur Folge haben kann. Das jeweils passende Modell zu wählen, hat entsprechend Auswirkungen auf die Klassifizierungsergebnisse. Spezialisierte Modelle wie *ClimateBERT-NetZero* bieten eine höhere Genauigkeit in spezifischen Kontexten, während allgemeine Modelle wie *GPT-4* flexibler sind, aber anfälliger für Fehlklassifikationen, insbesondere in hoch spezialisierten Kontexten. Dadurch wird die Sorgfaltspflicht bei der Auswahl des Modells in Abhängigkeit von Zielsetzung und Anwendungsfall deutlich, damit Effizienz und Präzision bestmöglich zueinander abgestimmt sind. Eine abschließende Klassifikation der 1.000 in diesem Abschnitt verwendeten US-amerikanischen Nachrichtenartikel zu freiwilligen Reduktionsversprechungen von CO_2-Emissionen, bei denen „Ja“ als zutreffend und „Nein“ als nicht zutreffend gilt, erfordert spezifisches Fachwissen und thematische Expertise. Daher ist eine Bewertung ohne entsprechendes Fachpersonal nicht möglich. Neben dem zeitlichen Aufwand pro Artikel stellt jedoch auch der finanzielle Aspekt eine zentrale Herausforderung dar. Im direkten Vergleich könnte ein KI-Modell (z. B. ein) eine effiziente Alternative darstellen, speziell wenn es gelingt, die Fehlerrate des Modells unter einem Schwellenwert von etwa 20% zu halten. Dies würde

statistisch signifikante Klassifizierungen ermöglichen und das Modell als unterstützendes Werkzeug in einer Vorauswahl etablieren.

Dieses Kapitel befasst sich entsprechend mit der Analyse und den Ergebnissen dreier LLMs, welche denselben Datensatz von 1.000 US-amerikanischen Nachrichtenartikel als Basis haben, um mögliche Unterschiede aber auch ihre jeweilige Nutzbarkeit herauszuarbeiten.

4.1 Aufbau

Der folgende schematische Prozessablauf in Abbildung 4.1 veranschaulicht den angewandten Modellablauf, der als Grundlage für die anschließende Ergebnisanalyse dient. Die Modelle folgen hierbei einem iterativen Klassifikationsprozess, der exemplarisch in Abbildung A.4 im Anhang dargestellt ist.

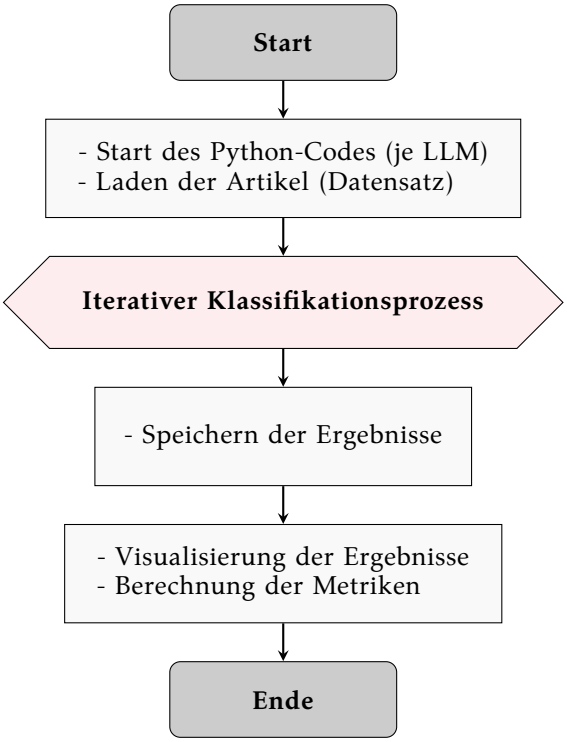

Abbildung 4.1: Schematischer Prozessablauf der LLM-Analyse

Zunächst werden die unterschiedlichen Modelle mit der Programmiersprache Python aufgesetzt. Daraufhin wird der Datensatz, der die 1.000 US-amerikanische Wirtschaftsnachrichtenartikel beinhaltet, eingelesen und zur iterativen Klassifizierung an das entsprechende Modell weitergeleitet. Die Resultate werden anschließend in externen Dateien abgespeichert, die für eine detailliertere grafische Auswertung und für die Berechnung der Modellmetriken verwendet werden. Die beschriebenen Abbildungen sowie weiterführende Erkenntnisse zu den Modellergebnissen sind im nachfolgenden Abschnitt 4.2 *Experimente* aufgeführt.

4.2 Experimente

Dieser Abschnitt führt die Modellergebnisse der LLMs *GPT-4*, *Gemma 2*, *Llama 3.1* und *ClimateBERT-NetZero* auf, welche unter anderem mithilfe der Python-*Quellcodes* A.1, A.2 und A.3 generiert worden. Aufgrund der Verschiedenheit der LLMs, werden interessante Unterschiede bei den Modellergebnissen zu erwarten sein. Diese Unterschiede werden anschließend im Teil 4.3 der *Auswertung* detailliert analysiert.

GPT-4

Die Klassifizierungsergebnisse des *GPT-4* Modells sind der nachfolgenden Abbildung 4.2 zu entnehmen. Die Abszisse zeigt die Zeit t in Jahren von 2005 bis 2023, während die Ordinate die prozentuale Verteilung der positiv klassifizierten Nachrichtenartikel auflistet. Die positiven *GPT-4*-Klassifizierungen werden durch die grünen Balken dargestellt. Als zusätzliche Verifizierungsebene sind in Orange die jährlichen Prozentsätze der positiven Experten Klassifizierungen aufgeführt. Die beiden Klassifizierungen haben die Form einer binären Klassifikationsmenge $L = \{0, 1\}$, wobei 0 für eine negative und 1 für eine positive Klassifikation steht.

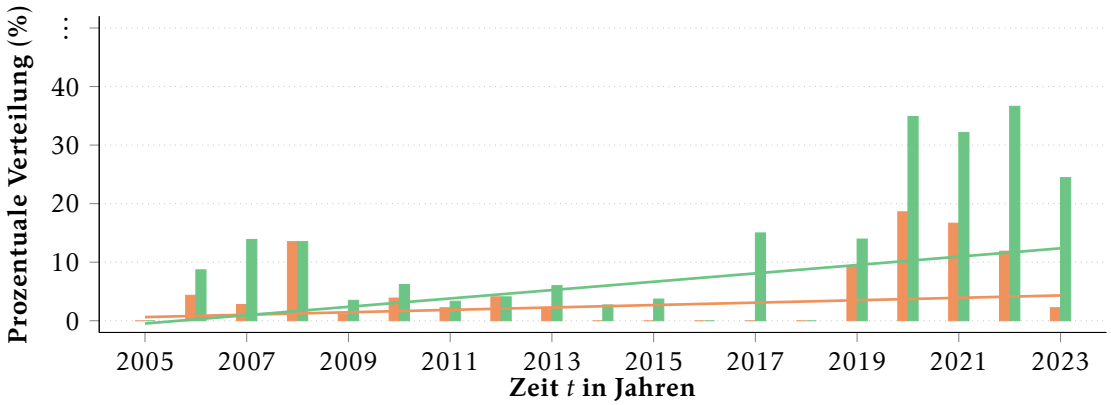

Abbildung 4.2: Prozentuale Verteilung positiver Artikel-Klassifikationen gemäß den Kategorien *Human-Coder* und *GPT-4*, die ein *neues* unternehmerisches Dekarbonisierungsversprechen von Treibhausgasemissionen repräsentieren.

Aufgrund der subjektiven Natur menschlicher Einschätzungen können diese Ergebnisse zwischen den Probanden erheblich variieren. Menschen neigen dazu, beim Interpretieren Kontext, implizite Hinweise, Nuancen und persönliche Erfahrungen einzubeziehen, was zu unterschiedlichen Klassifizierungen führen kann, auch wenn die Ausgangsvorgaben (gemäß der Klassifizierungsanleitung A.6) identisch sind. Ein häufig genutztes Maß zur Bewertung der *Intercoder-Reliabilität* ist Cohens κ, eine Korrelationsstatistik, die die Übereinstimmung

zwischen zwei unabhängig vorgenommenen Klassifikationen misst. (vgl. Grimmer, Roberts und Stewart (2022)). Die Einbeziehung des κ-Wertes führt aber zu weit in die Tiefe für diese Ausarbeitung. Im Gegensatz dazu, bewerten LLMs wie zum Beispiel *GPT-4*, solche Aufgaben auf Grundlage ihrer Basismodelle, vortrainierten Daten und auf ggf. bereits verarbeiteten Anfragen. Sie verwenden klare Formalien und explizite Sprachmuster, um Kontextabweichungen und implizite Bedeutungen zu minimieren, was hingegen jedoch zu systematischen Missklassifizierungen führen kann. Diese Ambiguität erfordert eine kritische Betrachtung der Ergebnisse, besonders da die objektiv richtigen Klassifikationen der Nachrichtenartikel unbekannt sind. Unterschiede in den Ergebnissen könnten eher von der zugrundeliegenden Methodik und den Trainingsdaten herrühren. Es ist somit entscheidend, dass das den Kontext der Thematik grundlegend versteht und darauf aufbauend sinnvoll arbeiten kann. Ein einfacher Basistest, in dem das nach seinem Wissen zu der Thematik der Emissionsreduktion befragt wird, kann als vorab Indikator dienlich sein, um das Modell auf seine Tauglichkeit zu überprüfen. Schlägt dies signifikant fehl, ist das Modell ungeeignet für weitere Analysen.

Bei Betrachtung der Abbildung 4.2 ist auffällig, wie verschieden die prozentualen Klassifizierungsausprägungen über die Zeit sind. Die positive Steigung der grünen Trendlinie des *GPT-4*-Modells und die jährlich ähnlichen Ausprägungen positiver Klassifikationen in Abbildung 4.2 legen nahe, dass das *GPT-4*-Modell den spezifisch semantischen Kontext erfassen kann und eine verwertbare Gesamtklassifizierung vornimmt. Im Gegensatz zu der grünen Trendlinie verläuft die orangene Trendlinie der Experten-Klassifikationen flacher, was auf eine konservativere Einschätzung positiver Artikel schließen lässt. Eine detailliertere Ergebnisanalyse folgt in Abschnitt 4.3.

Abbildung 4.2 zeigt jedoch keine Klassifikationsübereinstimmungen des Modells mit den Experten Klassifikationen auf. Es werden lediglich die prozentualen Verteilungen der positiven jährlichen Klassifikationen aufgetragen. Um eine Aussage über Übereinstimmungen und Unterschiede von Klassifikationsentscheidungen treffen zu können, wird anschließend im Abschnitt 4.3 *Auswertung*, noch beispielhaft eine Wahrheitsmatrix (Konfusionsmatrix) zu dem *GPT-4* Modell aufgestellt.

Gemma 2

Die grafische Auswertung der Klassifizierungsergebnisse des *Gemma 2*-Modells sind der folgenden Abbildung 4.3 zu entnehmen. Diese Modellergebnisse sind durch die Modellklassifikationen entsprechen ebenfalls der Form einer binären Lösungsmenge $L = \{0, 1\}$. In der Abbildung ist die prozentuale Verteilung der positiven klassifizierten Nachrichtenartikel in Abhängigkeit der erschienenen Jahre dargestellt. Hierbei ist die Zeit t in Jahren, beginnend von 2005 bis 2023 auf der

Abszisse aufgetragen und die prozentuale Verteilung der positiven klassifizierten Nachrichtenartikel auf der Ordinate aufgeführt. Die grünen Balken repräsentieren die jährlichen Prozentsätze der positiven *Gemma 2*-Klassifizierungen.

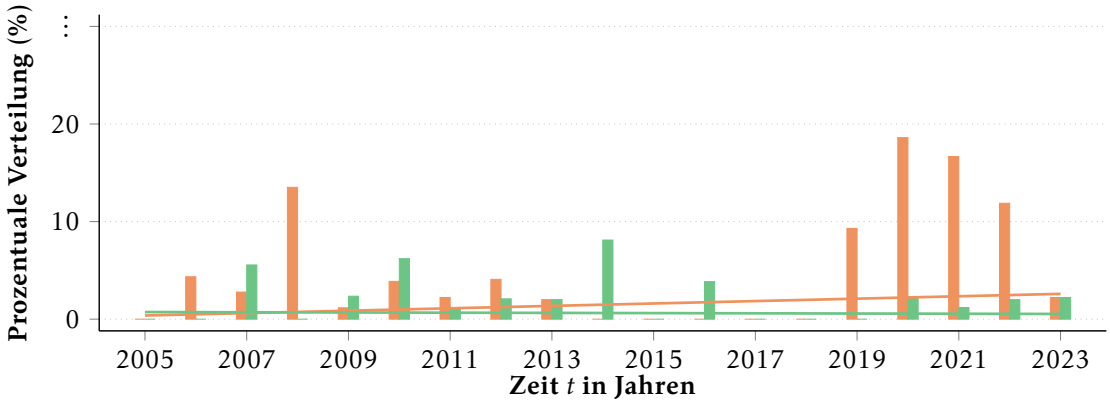

Abbildung 4.3: Prozentuale Verteilung positiver Artikel-Klassifikationen gemäß den Kategorien *Human-Coder* und *Gemma 2*, die ein *neues* unternehmerisches Dekarbonisierungsversprechen von Treibhausgasemissionen repräsentieren.

Abbildung 4.3 zeigt deutlich, wie ausgeprägt die Klassifizierungsdifferenzen im Laufe der Zeit sind, insbesondere bei den markanten Spitzen der Expertenklassifikationen in den Jahren 2008 sowie 2019 bis 2022. Die Steigung der grünen Trendlinie von *Gemma 2*, basierend auf den positiven Modellklassifizierungen, scheint nicht erkennbar positiv zu sein. Dies deutet darauf hin, dass das *Gemma 2*-Modell Schwierigkeiten hat, den spezifischen Kontext zu erfassen und nur selten positive Klassifizierungen vornimmt. Daher werden Unternehmensversprechungen und -Zielsetzungen zur Emissionsreduktion häufig als strikt negativ eingestuft. Abschnitt 4.3 enthält eine detailliertere Analyse der Ergebnisse.

Llama 3.1

In der nachfolgenden Abbildung 4.4 sind die Klassifizierungsergebnisse des *Llama 3.1* Modells zu entnehmen. Die Klassifizierungsergebnisse des *Llama 3.1*-Modells entsprechen ebenfalls der Form einer binären Lösungsmenge von $L = \{0, 1\}$. In der Abbildung ist die Zeit t in Jahren, beginnend von 2005 bis 2023 auf der Abszisse aufgetragen und die prozentuale Verteilung der positiven klassifizierten Nachrichtenartikel auf der Ordinate aufgeführt. Die grünen Balken repräsentieren die jährlichen Prozentsätze der positiven *Llama 3.1*-Klassifizierungen.

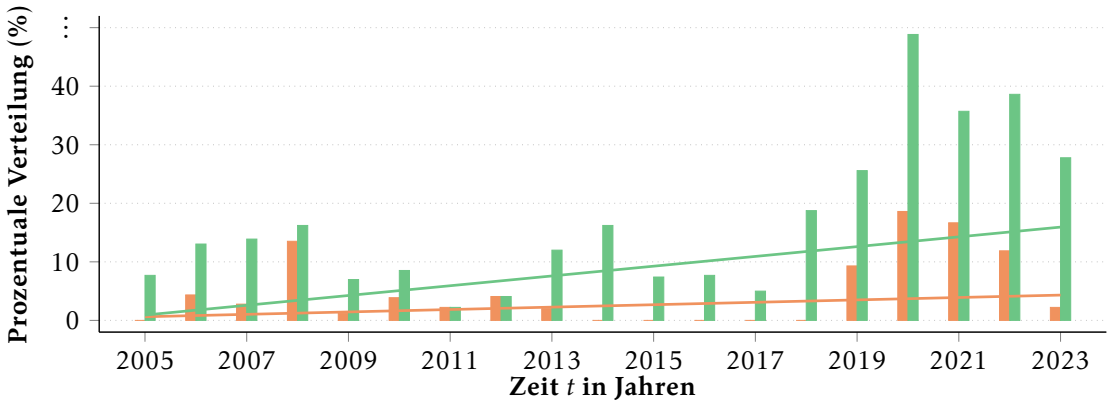

Abbildung 4.4: Prozentuale Verteilung positiver Artikel-Klassifikationen gemäß den Kategorien *Human-Coder* und *Llama 3.1*, die ein *neues* unternehmerisches Dekarbonisierungsversprechen von Treibhausgasemissionen repräsentieren.

Bei Betrachtung der Abbildung 4.4 ist auffällig, wie auch hier die Klassifizierungsdifferenzen über die Zeit Ausprägung sind. Die positive Steigung der grünen Trendlinie des *Llama 3.1*-Modells und die jährlich ähnlichen Ausprägungen der prozentualen Verteilungen positiver Klassifikationen in Abbildung 4.4 legen nahe, dass das *Llama 3.1*-Modell den spezifisch semantischen Kontext erfassen kann und eine verwertbare Gesamtklassifizierung vornimmt. Im Gegensatz zu der grünen Trendlinie verläuft die orangene Trendlinie der Experten-Klassifikationen flacher, was auf eine konservativere Einschätzung positiver Artikel schließen lässt. Eine vergleichbar höhere Klassifizierungshäufigkeit konnte zuvor bei den Ergebnissen des *GPT-4*-Modells ebenfalls festgestellt werden. Eine detailliertere Ergebnisanalyse folgt in Abschnitt 4.3.

ClimateBERT-NetZero

Im Gegensatz zu den anderen LLMs zeigt das *ClimateBERT-NetZero*-Modell seine Stärke darin, eine allgemeine Vorselektion von Kontext relevanten Artikeln zu treffen. Dieses liegt daran, da das Modell nicht miteinbezieht, ob es sich bei dem Artikel um eine *neue* Emissionsreduktionsbekundung des Unternehmens handelt oder ob es sich um eine allgemeine und wiederholte Bekanntmachung handelt. Entsprechend lassen sich die Klassifikationsergebnisse nicht mit denen der Experten vergleichen. Der nachfolgenden Abbildung 4.5 ist zu entnehmen, wie das *ClimateBERT-NetZero*-Modell von Schimanski u. a. (2023), selbigen Datensatz von 1.000 US-amerikanischen Nachrichtenartikeln, gemäß der Kategorien der Lösungsmenge $L = \{\text{Ohne Bezug}, \text{Reduktion}, \text{Net-Zero}\}$, klassifiziert hat. Auch hier ist die Zeit t in Jahren, beginnend von 2005 bis 2023 auf der Abszisse aufgetragen und die prozentuale Verteilung der positiven klassifizierten Nachrichtenartikel auf der Ordinate aufgeführt. Die grünen Balken repräsentieren die

jährlichen Prozentsätze der positiv Net-Zero Klassifizierungen. Wobei die blauen Balken die jährlichen Prozentsätze der positiven Reduktionsversprechungen kennzeichnen.

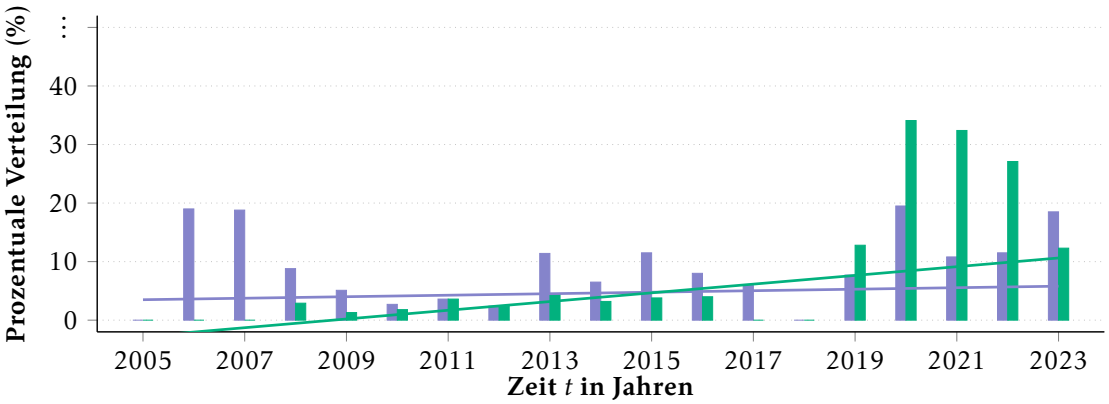

Abbildung 4.5: Prozentuale Verteilung positiver Artikel-Klassifikationen gemäß den Kategorien *Reduktion* und *Net-Zero*, die unternehmerische Dekarbonisierungsversprechen von Treibhausgasemissionen repräsentieren.

Das *ClimateBERT-NetZero* Modell hat insgesamt 82, 60% (826 Artikel) der Artikel als *Ohne Bezug*, 8, 20% (82 Artikel) als *Reduktion* und 9, 20% (92 Artikel) als *Net-Zero* klassifiziert. Es wird ersichtlich, dass im Verlaufe der Zeit vermehrt Net-Zero Klassifikationen auftreten. Diese Beobachtungen und Ergebnisse gleichen sich mit den Aussagen von Ardia u. a. (2023). Über die Zeit betrachtet, scheinen vermehrt Unternehmen sich einer freiwilligen Emissionsreduktion zu verschreiben und sich ebenfalls viele Unternehmen einer positiven Net-Zero Bewegung anzustreben. Es gilt jedoch zu bedenken, dass hier ein begrenzter Datensatz genutzt wurde und dies lediglich einem Ausschnitt der Realität entsprecht. Im direkten Vergleich zu *GPT-4* demonstriert *ClimateBERT-NetZero* in diesem spezifischen Kontext, dass spezialisierte, optimierte und effiziente Transformermodelle ohne ständig wachsende Modellstrukturen wettbewerbsfähig sein können. ClimateBERT zeigt beispielhaft, wie Open-Source-Ansätze schnelle und gleichzeitig leistungsstarke Klassifikationsergebnisse liefern können, indem es Tokenisierungs- und Trainingsprozesseffiziente Technologien nutzt (vgl. Sanh (2019)).

Vorab wurden 96 der 1.000 Nachrichtenartikel durch das Basismodell *ClimateBERT* (distilroberta-base-climate-detector) als kontextspezifisch irrelevant deklariert und aussortiert. Diese Artikel zählen sich somit in die Gruppe der als *Ohne Bezug* zu deklarierenden Artikel. Anschließend wurden die verbliebenen 904 Artikel weiter an das spezialisierte *ClimateBERT* (netzero-reduction) Modell für die tiefergehende Klassifizierung weitergeleitet.

Sowohl das *ClimateBERT-NetZero*-Modell, als auch das *GPT-4*-Modell, zeigen über die Zeitspanne der letzten fünf Jahre des Datensatzes ein überdurchschnitt-

lich erhöhtes Aufkommen von Unternehmen auf, die vermehrt ehrgeizige Emissionsreduktionsziele beziehungsweise Net-Zero Ziele angekündigt haben. Dies könnte direkt mit der wachsenden Besorgnis über den Klimawandel in Zusammenhang stehen (vgl. Ardia u. a. (2023)). Eine genauere Untersuchung der positiv klassifizierten Artikel zeigt, dass vor allem große und „braune“ Unternehmen (solche mit hohen Emissionsintensitäten) freiwillig grüne Verpflichtungen eingehen, sowohl innerhalb ihrer Branchen als auch branchenübergreifend. Weil diese großen, „braunen“ Unternehmen in den USA für den Übergang zu einer kohlenstoffarmen Wirtschaft besonders relevant sind, haben ihre Reduktionsverpflichtungen eine entsprechend große Bedeutung (vgl. Bauer u. a. (Nov 2024) und Acharya, Engle und Wang (2025)).

Die zuvor gewonnenen ersten Erkenntnisse und weiterführenden Überlegungen führen zu einer zentralen Frage: Können speziell trainierte Open-Source-Modelle wie *ClimateBERT-NetZero*, die mit kontextspezifisch passenden Textdaten trainiert wurden, oder allgemein vortrainierte Modelle wie *Gemma 2* oder *Llama 3.1*, einem nicht Open-Source-Modell wie *GPT-4*, das größer, breiter vortrainiert und allgemein vielseitiger ist, ebenbürtig oder sogar schon überlegen sein und ebenfalls zu sinnvollen Erkenntnissen beitragen? Dies kann als Überblick über die zuvor formulierten Forschungsfragen verstanden werden und wird durch die nachfolgende Analyse der Experimente beantwortet.

4.3 Auswertung

Um eine Vergleichbarkeit der Modelle zu ermöglichen, bietet es sich an, die Trendlinien über die Zeit zu betrachten. Hierzu wird zunächst die Steigung der jeweiligen Trendlinie mithilfe folgender Steigungsgleichung m errechnet.

$$m = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N} (x_i - \overline{x})^2}$$

Wobei x als unabhängige Variable für die Zeit in Jahren und y als abhängige Variable für den prozentualen Wert des jeweiligen Klassifikationsaufkommens in dem Jahr steht. $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ und $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ entsprechen jeweils den mittleren Stichprobenmengen. Dabei zählen die Summen, beginnend bei $i = 1, ..., N$, die Anzahl an Nachrichtenartikeln entsprechend auf. Hierbei misst der Zähler die Kovarianz und der Nenner die Varianz von x (vgl. James u. a. (2023); S. 71-72).

Die Analyse der Trendlinien zeigt, dass die Steigung des *GPT-4* Modells bei $m_{\text{GPT}} = 1,319$ liegt, während das *ClimateBERT-NetZero*-Modell eine Steigung

von $m_{\text{CBERT}} = 1.375$ aufweist. Hierbei wurden die Steigungen der Reduktion und Net-Zero vereint, um eine Vergleichbarkeit herzustellen. Die Steigungsdifferenz der *GPT-4* und ClimateBERT NetZero Trendlinien beträgt $\Delta m_1 = m_{\text{CBERT}} - m_{\text{GPT}} = 0,056$. Somit klassifiziert das ClimateBERT-NetZero-Modell durchschnittlich pro Jahr $\frac{\Delta m_1}{m_{\text{GPT}}} \cdot 100 = 4,25\%$ Nachrichtenartikel mehr positiv, als das *GPT-4*-Modell. In direkter Relation, entspricht dies einer geringen Änderung ($|\Delta m_1| < 10\%$) und deutet darauf hin, dass beide Modelle, *GPT-4* und *ClimateBERT NetZero*, ähnlich auf den semantischen Kontext der Nachrichtenartikel über die Zeit reagieren und deren Klassifizierungsergebnisse über die Jahre hinweg konsistent sind. Obwohl die Auswertungsdauer des *GPT-4*-Modells nicht bekannt ist, weist die Steigungsdifferenz von lediglich $4,25\%$ darauf hin, dass das *ClimateBERT-NetZero*-Modell eine interessante Open-Source-Alternative darstellen könnte. Besonders bemerkenswert ist die kurze und effiziente Klassifizierungszeit des *ClimateBERT-NetZero*-Modells, die neben einer potenziellen Kostenersparnis auch dessen Konkurrenzfähigkeit unterstreicht. Eine Differenz von unter 10% in den Trendlinien der Modelle kann, abhängig vom Anwendungsfall, als gering, aber dennoch relevant interpretiert werden. Während bei rein praktischen Anwendungen eine solche Abweichung als unproblematisch gelten kann, können beispielsweise politische, ökologische oder wirtschaftliche Analysen bereits durch kleinere Unterschiede signifikant beeinflusst werden. Um die statistische Signifikanz der Steigungsdifferenz besser einordnen zu können, wäre es möglich ein t-Test vorzunehmen, um zu prüfen, ob die beobachteten Unterschiede zufallsbedingte oder tatsächliche Unterschiede sind und auf inhärente Modellunterschiede hinweisen. Bei diesen Direktvergleichen der Trendsteigungen greift aber erstere Interpreation und kann entsprechend als gering und positiv angesehen werden. Auch wenn *ClimateBERT-NetZero* im direkten Vergleich zu *GPT-4* einen ähnlichen Trend abbildet, könnte es dennoch an dem notwendigen semantischen Verständnis fehlen, das für eine tiefere Analyse und eine präzisere Bewertung von Veränderungen in der Unternehmensentwicklung erforderlich wäre. Dies liegt an der spezifischeren und eingeschränkteren Basisdatenlage und könnte in Zusammenhang mit der allgemeineren Aufgabenlösung ein weiterer Erklärungsansatz für die Unterschiede in den Abbildungen dienen.

Ähnlich wie das *GPT-4*-Modell kann auch das *Llama-3.1*-Modell durch präzise Anweisungen, die während des Klassifizierungsprozesses übermittelt werden, darauf ausgerichtet werden, *neue* Emissionsreduktionen gemäß den Vorgaben aus A.6 optimal zu klassifizieren. Dadurch können die Resultate und Trendlinien des *GPT-4*- und *Llama 3.1*-Modells genauer miteinander verglichen und bewertet werden. Die *Llama-3.1*-Trendlinie hat eine Steigung von $m_{\text{LLAMA}} = 1,542$ und die Differenz beider Geradensteigungen beträgt $\Delta m_2 = m_{\text{LLAMA}} - m_{\text{GPT}} = 0,223$. Das Llama-3.1-Modell weist eine jährliche durchschnittliche positive Klassifizierung

von $\frac{\Delta m_2}{m_{\text{GPT}}} \cdot 100 = 16,91\%$ mehr Artikeln im Vergleich zum *GPT-4*-Modell auf. In direkter Relation, entspricht dies einer signifikanten Differenz im Verlauf der positiven Steigungen der Trendlinien, da ($|\Delta m_2| > 10\%$) und deutet darauf hin, dass beide Modelle, *GPT-4* und *Llama 3.1*, verschieden auf den semantischen Kontext der Nachrichtenartikel über die Zeit reagieren und deren Klassifizierungsergebnisse aber über die Jahre hinweg konsistent sind.

Die Betrachtung der Trendlinie des *Gemma 2* Modells zeigt mit $m_{\text{GEMMA}} = -0,033$ hingegen eine negative Steigung auf, welche im Widersprich zu den vorherigen Erkenntnissen steht. Im Vergleich zu Ardia u. a. (2023), sollte der Trend dieser Linie positiv verlaufen. Auch die direkte Differenz zu der *GPT-4*-Trendlinie ist auffällig groß, mit $\Delta m_3 = m_{\text{GEMMA}} - m_{\text{GPT}} = -1,352$ bzw. einer prozentualen jährlichen Klassifikationsdifferenz gegenüber der GPT-Trendlinie von $\frac{\Delta m_3}{m_{\text{GPT}}} \cdot 100 = -102,53\%$. In direkter Relation, entspricht dies einer signifikanten Änderung ($|\Delta m_3| > 10\%$) und deutet zusammenfassend darauf hin, dass das *Gemma 2*-Modell den semantischen Kontext nicht erfassen kann und seine Aufgabe nicht wie gewünscht erfüllen kann. Eine umfangreichere Bewertung und Vergleichbarkeit dieser Schlussfolgerungen kann erst in Kombination mithilfe der aus dem Abschnitt 3.7 aufgestellten Metriken getroffen werden. Tabelle 4.1 listet vorab die bis hierhin aufgeführten Daten und gewonnen Erkenntnisse auf.

Tabelle 4.1: Ergebnisübersicht verwendeter LLMs für den Klassifizierungsprozess.

Modell-Name	**Kontextspezifisch vortrainiert**	**Laufzeit**	**Trend**
GPT-4 (*gpt-4-0613*) *OpenAI*, Anzahl der Token Unbekannt	Nein	Unbekannt	1,319
Gemma 2 (*ff02c3702f32*) *Google AI*, 9B Token	Nein	1:37:10	-0,033
Llama 3.1 (*46e0c10c039e*) *Meta AI*, 8B Token	Nein	0:17:27	1,542
ClimateBERT (*netzero-reduction*) Forscherteam der *University of Cambridge*, 82M Token	Ja	0:00:28	1,375

Die in der vorangegangenen Tabelle 4.1 aufgeführten Laufzeiten der LLMs ermöglichen eine direkte Vergleichbarkeit mit einer menschlichen Klassifizierungsgeschwindigkeit. Nimmt man exemplarisch an, dass ein Mensch im Durchschnitt 10 Minuten pro Klassifizierung eines Nachrichtenartikels benötigt, würde die Bearbeitung der Menge an Artikeln rund 10.000 Minuten (in etwa 167 Stun-

den) in Anspruch nehmen. Im Gegensatz dazu benötigt beispielsweise das LLM-Modell *Llama 3.1* rund 17 Minuten für denselben Klassifizierungsprozess. Dies führt zu einem signifikanten Effizienzgewinn, da der Klassifizierungsprozess erheblich beschleunigt wird. Somit absolviert *Llama 3.1* den hier exemplarisch aufgeführten Klassifizierungsprozess vom Faktor 573-mal schneller als ein Mensch.

Die Betrachtung der Geschwindigkeit allein reicht nicht aus, um die Effektivität von LLMs anhand dieses gewählten Beispieles zu beurteilen. LLMs müssen ebenfalls in Verbindung mit ihrer semantischen Genauigkeit bewertet werden. Wenn *Llama 3.1* schneller, aber nur geringfügig weniger genau als der Mensch arbeitet, könnte das LLM zunächst einmal als praktikables Werkzeug betrachtet werden, insbesondere bei großen Datensätzen. Sollte sich zudem herausstellen, dass die semantische Tiefe oder die Erfassung subtiler Kontexte wie impliziter Dekarbonisierungsversprechen mangelhaft ist, deutet dies auf Grenzen hin, die ggf. mithilfe von dem Kontext speziell vortrainierten Modellen wie ClimateBERT-NetZero entsprechend entgegengetreten werden kann. Um entsprechend die LLM-Effektivität hinsichtlich ihrer semantischen Genauigkeit zu prüfen, soll folgendes exemplarisches Beispiel dienen.

		Human Coder		
		Negativ	**Positiv**	Total
GPT-4	**Negativ**	842	15	857
	Positiv	100	43	143
	Total	942	58	1.000

Precision:	0,30	**Recall:**	0,74
F_1 score:	0,43	**Accuracy:**	0,89

Abbildung 4.6: Übersicht der GPT-4-Modellergebnisse in Form einer Konfusionsmatrix mit Metriken.

Abbildung 4.6 repräsentiert die Konfusionsmatrix, welche die quantitativen Metriken für die semantische Genauigkeit von *GPT-4*-Klassifikationen im Vergleich zu den Expertenklassifikationen für den Zeitraum zwischen den Jahren 2005 und 2023 darstellt. Die Konfusionsmatrix stellt in horizontaler Richtung die von *GPT-4* klassifizierten (geschätzten) Klassifikationen auf und in vertikaler Richtung die von den Experten klassifizierten (*wahren*) Klassifikationen auf. Durch die Gliederung in *falsch Negativ* (f_n), *falsch Positiv* (f_p), *richtig Negativ* (r_n) und *richtig Positiv* (r_p) Klassifikationen wird eine detaillierte Untersuchung der Klassifikationsgüte ermöglicht (vgl. Asim, Rehman und Shoaib (2017); S. 372). So veranschaulicht sie die semantisch kontextspezifische Leistung des Modells.

Die Metriken aus Abbildung 4.6 zeigen eine gemischte Modellgüte auf und erlauben es eine tiefergehende Bewertung der Eignung des Modells für die semantische

Analyse und Textklassifizierung. Mit einer Accuracy von 89% und einer Specificity von 89% zeigt das *GPT-4*-Modell eine vielversprechende Fähigkeit, korrekte Klassifikationen, insbesondere im negativen Fallbereich, vorzunehmen. Der Recall von 74% deutet darauf hin, dass das LMM relevante positive Klassifikationen gut erkennt und semantische Zusammenhänge erfassen kann. Jedoch weist die geringe Precision von 30% auf eine hohe Anzahl falsch-positiver Klassifikationen hin, was auf eine geringe Verlässlichkeit bei der Erkennung positiver Klassen hindeutet. Das niedrige F-Maß von 43% verdeutlicht das Ungleichgewicht zwischen Precision und Recall und zeigt, dass die Gesamtleistung des Modells suboptimal ist. Dies wiederum ist speziell für Anwendungen wichtig, bei denen die Präzision entscheidend ist. Zusammenfassend eignet sich das Modell nur bedingt für die hier gestellte Aufgabe und bedarf einer Verbesserung der Precision, um Praxistauglichkeit des LLMs zu erhöhen.

An diesem Punkt ist es wichtig hervorzuheben, dass die Klassifikationen der Experten hier als *Ground Truth* verwendet wurden, die jedoch nicht zwangsläufig einer „wahren" Klassifikation entsprechen müssen. Die Annahme der *Ground Truth* repräsentiert lediglich exemplarisch die tatsächlichen korrekten Klassifikationen, gegen die die Modellvorhersagen validiert wurden, um die Modellleistung anhand der Metriken zu messen. Diese Vorgehensweise dient primär der Veranschaulichung der methodischen Auswertung. Entsprechend sollten finalen Schlussfolgerungen als beispielhaft und nicht als abschließend betrachtet werden.

Als weiterer Vorteil von LLMs kann jedoch ihre konstante Leistungsfähigkeit angesehen werden. Menschen, die codieren, können aufgrund von Ermüdung, Ablenkung oder anderen äußeren Einflüssen Schwankungen in ihrer Leistung aufweisen. Im Gegensatz dazu liefern LLMs über längere Zeiträume hinweg konstante Ergebnisse, wodurch Inkonsistenzen im Klassifizierungsprozess vermieden werden. LLMs sind aufgrund ihrer Fähigkeit, bei wiederholten oder umfangreichen Aufgaben konstant und fehlerfrei zu arbeiten, wertvolle Hilfsmittel in der modernen Textanalyse. Hierbei wird deutlich, dass LLMs, wenn sie korrekt angewendet werden, den Klassifizierungsprozess effizienter machen und menschliche Inkonsistenzen verringern könnten. Dies hebt die Anwendbarkeit von LLMs in zahlreichen Anwendungsbereichen hervor, sofern die LLMs richtig ausgewählt und in passende Kontexte eingebunden werden.

Die Gesamtübersichten der Experimentauswertungen im Anhang (Abbildung A.5 und die Gesamtübersicht aller LLM-Konfusionsmatrizen und Metriken im Abschnitt A.5) verdeutlichen zusammenfassend, dass die Wahl des LLMs von einer Abwägung zwischen Geschwindigkeit, Kosten und der erforderlichen Präzision im spezifischen Kontext der Sentimentanalyse abhängt. *Llama 3.1*, *GPT-4*

und *ClimateBERT-NetZero* erweisen sich allesamt als geeignete Instrumente für tiefergehende Analysen.

Während *ClimateBERT-NetZero* aufgrund seiner Effizienz für schnelle, allgemeine klimaspezifische Voranalysen bevorzugt werden kann, sollte für präzisere und konsistentere Klassifikationen auf ein flexibleres Modell wie *Llama 3.1* zurückgegriffen werden. Im direkten Vergleich zu den Ergebnissen von *GPT-4* ist es notwendig, die finalen Klassifikationen auf ihre Kontextgenauigkeit zu überprüfen. Dies unterstreicht die Bedeutung der Modellauswahl im Hinblick auf die spezifischen Anforderungen der Aufgabe.

Ergebnisübersicht der LLM-Klassifizierungsmetriken

Die folgende tabellarische Ergebnisübersicht aller LLM-Klassifizierungsmetriken, stellt eine zusammenfassende Gesamtübersicht der LLM-Konfusionsmatrizen aus dem Anhang A.5 (A.6, A.7 und A.8) dar. Sie listet die unterschiedlichen semantischen Modellgüten auf, wobei alle Modellklassifikationen mithilfe der menschlichen Klassifikationen (Experten Klassifikationen) und der Annahme des *Ground Truth* erstellt worden, um entsprechende konsistente Metriken und resultierende Unterschiede erkennbar zu machen.

Tabelle 4.2: Ergebnisübersicht der LLM-Klassifizierungsmetriken.

Modell-Name	**Precision**	**Recall**	**F_1**	**Accuracy**
GPT-4 (*gpt-4-0613*)	0,30	0,74	0,43	0,89
Gemma 2 (*ff02c3702f32*)	0,00	0,00	0,00	0,92
Llama 3.1 (*46e0c10c039e*)	0,22	0,69	0,33	0,84
ClimateBERT (*netzero-reduction*)	–	–	–	–

Bei Betrachtung der Tabelle 4.2 fallen besonders die niedrigen Precision und Recall Werte des *Gemma 2*-Modells von 0, 00 auf. Diese Werte begründen sich daraus, dass das Modell keinen einzigen richtig Positiv (r_p) Fall identifizieren konnte. Infolgedessen ergibt sich ein F-Maß von ebenfalls 0, 00, da dieser das harmonische Mittel aus Precision und Recall darstellt. Die Accuracy des Modells erscheint, im Vergleich zu den anderen Modellen, auf den ersten Blick hoch. Dieser Wert begründet sich aber daraus, dass das Modell die meisten Artikel im Datensatz negativ (Klassifizierung: 0) klassifiziert hat. Da das Modell diesen Fall im Gesamtverhältnis fast immer korrekt klassifiziert, ergibt sich eine hohe Anzahl an richtig Negativ ($r_n = 917$) klassifizierten Fällen und somit eine hohe Treffergenauigkeit des Modells. Allerdings sollte dieser hohe Accuracy-Wert nicht über die schlechte Gesamtleistung hinwegtäuschen, da das Modell vollständig alle r_p Fälle verfehlt hat. Dieses Ergebnis verdeutlicht, dass eine hohe Accuracy nicht zwangsläufig auf eine ausgewogene oder zufriedenstellende Modellleistung hindeutet, was insbesondere bei unbalancierten Datensätzen auftreten kann. Das *Gemma 2*-Modell erweist

sich in diesem speziellen Anwendungsfall nicht nur als unzureichend leistungsfähig, sondern weist auch erhebliche Mängel bei der Erfassung des semantischen Kontexts auf.

Bei dem *ClimateBERT-NetZero*-Modell hingegen gibt es keine abschließenden Metriken, da im gegensatz zu den anderen drei Modellen, die Ergebnisklassifikationen gemäß den folgenden eigens übersetzten Definitionen nach Schimanski u. a. (2023) einer Multi-Klassen-Klassifikationsart entsprechen und deklariert sind:

- *Reduktionsziele* sind Aussagen, die sich auf eine absolute oder relative Verringerung von Emissionen beziehen, oft begleitet von einem Basisjahr, mit dem das Reduktionsziel verglichen wird.
- *Netto-Null-Ziele* stellen einen Sonderfall von Reduktionszielen dar, bei dem eine Institution erklärt, ihre Emissionsbilanz bis zu einem bestimmten Jahr auf keine zusätzlichen Nettoemissionen zu reduzieren.
- *Wenn beide Zielarten im Text erscheinen*, ist der Hauptfokus des Textes entscheidend. Zum Beispiel dienen die meisten Reduktionsziele als Zwischenschritte für das endgültige Ziel von Netto-Null. Daher liegt der Schwerpunkt auf Netto-Null.

Demzufolge wird in dem Modell nicht der Fokus auf eine *neue* Emissionsreduktion gelegt, sondern der generelle Fall einer steigenden Reduktion oder Net-Zero Emissionsreduktion klassifiziert. Hierdurch ist es nicht möglich auf die Metriken der Klassifikationen zu schließen, wie es bei den anderen Modellen der Fall war. Dennoch kann gerade dieses speziell vortrainierte Modell, mit seiner schnellen Laufzeit von *28 Sekunden* für die *1.000 US-amerikanischen Nachrichtenartikeln*, sehr nützlich sein, um eine genauere Vorauswahl relevanter Artikel vorzunehmen und anschließend diese relevanten Artikel auf *neue* Emissionsreduktionen zu überprüfen. Die in dieser Arbeit erzielten Modellergebnisse weisen in ihren Kurvenverläufen deutliche Parallelen zu den Resultaten des wissenschaftlichen Berichts der ClimateBERT-NetZero-Modellentwickler auf (vgl. Schimanski u. a. (2023), S. 5, Abbildung 1). Diese Ergebnisse bekräftigen die Anwendbarkeit des Modells für den vorliegenden Datensatz und den spezifischen Kontext.

5 Schlussbetrachtung

5.1 Zusammenfassung

Zusammenfassend lässt sich festhalten, dass mithilfe der in dieser Arbeit angewandten Experimente relevante Erkenntnisse gewonnen werden konnten und Schlussfolgerungen ermöglicht haben, die bewiesen, dass Identifizierungen von freiwilligen Dekarbonisierungsversprechen von Firmen mittels Large-Language-Modellen zielführend umgesetzt werden konnte. Modelle wie *ClimateBERT-NetZero* haben sich bei klimabezogenen Themen durch spezialisierte Trainingsdaten als effektiv erwiesen, während *GPT-4* und *Llama 3.1* mit ihrer Flexibilität und semantischen Genauigkeit überzeugten (vgl. Achiam u. a. 2023). Allerdings gibt es Einschränkungen bei Erkennung tiefergehender subtiler Kontexte oder impliziter Verpflichtungen und bedarf einer genaueren Gegen-Verifizierung, mithilfe von Expertenklassifikationen, um eine gewisse Modellgüte zu gewährleisten. Ein wesentlicher Aspekt, der nicht außer Acht gelassen werden sollte, ist, dass die Ergebnisse des Modells nicht ausschließlich auf den eingelesenen Daten beruhen. Auch die Kontrolle darüber, ob das Modell im iterativen Klassifikationsprozess für jede Artikelklassifikation eine neue Instanz generiert oder kontinuierlich auf zuvor gesammelten Daten basiert und somit von bereits erlerntem Wissen beeinflusst wird, ist entscheidend. Dies ist nicht nur für die in dieser Studie durchgeführten Experimente relevant, sondern stellt einen grundlegenden Kontrollfaktor dar, um die gewünschte Modellantwort und deren Qualität gezielt abrufen zu können. Um fundierte Schlussfolgerungen aus den Gesamtergebnissen in Bezug auf die eingelesenen Daten ziehen und gegebenenfalls gezielte Anpassungen vornehmen zu können, ist es unerlässlich, die genauen Abläufe eines Modells tiefgehend zu verstehen. In der Summe lässt sich festhalten, dass LLMs ein effizientes Hilfsmittel sind, das menschliche Analysen sinnvoll ergänzen kann, aber nicht vollständig ersetzt. Mit angemessener Implementierung und passender Aufgabenzuweisung der LLMs, ließ sich zeigen, dass die Effizienz signifikant gesteigert werden konnte und in Kombination mit Fachwissen (Kontrollklassifikationen) und weiterer Auswertungen interessante Anwendungsbereiche sich in nahezu alle möglichen Fachgebieten eröffnen.

5.2 Ausblick

Bei weiterführenden Forschungen im Bereich der Anwendung und Weiterentwicklung von Sprachmodellen ist jedoch ein ausgewogenes Verhältnis zwischen Rechenaufwand und Ergebnisqualität essenziell, da größere Modelle nicht zwangsläufig bessere Ergebnisse liefern müssen. Eine detaillierte Analyse der Ressour-

cennutzung könnte dazu beitragen, effizientere Modelle zu entwickeln, die sowohl präzise als auch ressourcenschonend arbeiten. Ein weiterer wichtiger Aspekt kann die Integration zusätzlicher Sprachmodelle sein, wie etwa Mixtral von Mistral, die eine noch größere sprachliche Abdeckung bieten. Insbesondere der Einbezug von Modellen, die den chinesischen Markt abdecken können, können sehr Interesse sein, da China als weltweit größter CO_2-Emittent eine zentrale Rolle in globalen Klimastrategien spielt. Eine solche Erweiterung würde den Zugang zu globalen Nachrichtenquellen ermöglichen und die Tiefe sowie die Vielfalt zukünftiger Analysen steigern. Zusätzlich könnten spezifische Anpassungen an kontextspezifischen Modellen, wie *ClimateBERT-NetZero*, untersucht werden. Beispielsweise wäre es denkbar, das Modell so zu modifizieren, dass ausschließlich neue Emissionsreduktionen oder Netto-Null-Versprechen als positiv klassifiziert werden. Eine solche Feinjustierung würde die Stärke vortrainierter Modelle besser ausnutzen und die Präzision bei der Identifikation von relevanten Artikeln erhöhen. Auch die Qualität der Datengrundlage bietet Raum für Verbesserungen. Ein optimierter Web-Scraping-Prozess oder eine umfassende Nachbearbeitung der gesammelten Daten könnte dazu beitragen, eine klarere und konsistentere Datenbasis zu schaffen. Dies würde nicht nur die Klassifikationsergebnisse, sondern auch die Aussagekraft der Analysen insgesamt steigern. Schließlich könnte die Weiterentwicklung der analysierten Methoden praktische Anwendungen fördern, wie etwa die systematische Überprüfung der Einhaltung von Netto-Null-Zusagen oder die Identifikation von Diskrepanzen zwischen kommunizierten Zielen und tatsächlichen Maßnahmen. Solche Anwendungen könnten nicht nur Unternehmen und politische Akteure unterstützen, sondern auch einen Beitrag zur globalen Klimapolitik leisten.

Insgesamt zeigt diese Arbeit auf, dass die Kombination aus innovativen Sprachmodellen, optimierten Datenstrategien und zielgerichteten Anpassungen eine Schlüsselrolle bei der Analyse und Bewertung von klimapolitischen Maßnahmen spielt. Abschließend weisen die hier vorgestellten Perspektiven den Weg für zukünftige Forschungen und praktische Anwendungen, die sowohl wissenschaftlich als auch gesellschaftlich von Relevanz sind.

Literatur

Acharya, Viral V., Robert F. Engle und Olivier Wang (2025). *Strategic commitments to decarbonize: The role of large firms, common ownership, and governments*. Techn. Ber. Working Paper. National Bureau of Economic Research. DOI: https://dx.doi.org/10.3386/w33335. Letzter Abruf am 19. Januar 2025 um 18:16 Uhr.

Ahmad, Khaleeq u. a. (2023). „Greenhouse gas emissions and corporate social responsibility in USA: A comprehensive study using dynamic panel model“. In: *Heliyon* 9.3. DOI: http://doi.org/10.1016/j.heliyon.2023.e13979. Letzter Abruf am 19. Januar 2025 um 18:30 Uhr.

Ardia, David u. a. (2023). „Climate Change Concerns and the Performance of Green vs. Brown Stocks“. In: *Management Science* 69.12, S. 7607–7632. DOI: https://doi.org/10.1287/mnsc.2022.4636. Letzter Abruf am 19. Januar 2025 um 18:45 Uhr.

Asim, Muhammad Nabeel, Abdur Rehman und Umar Shoaib (2017). „Accuracy Based Feature Ranking Metric for Multi-Label Text Classification“. In: *International Journal of Advanced Computer Science and Applications* 8.10. DOI: https://doi.org/10.14569/IJACSA.2017.081048. Letzter Abruf am 19. Januar 2025 um 19:25 Uhr.

Bauer, M. u. a. (Nov 2024). „Corporate Green Pledges“. SSRN Working Paper. DOI: https://dx.doi.org/10.2139/ssrn.5027881. Letzter Abruf am 19. Januar 2025 um 18:09 Uhr.

Döbel, Inga u. a. (2018). „Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung“. In: *Fraunhofer-Gesellschaft, München*. URL: https://www.bigdata-ai.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf. Letzter Abruf am 19. Januar 2025 um 19:20 Uhr.

Grimmer, Justin, Margaret E Roberts und Brandon M Stewart (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Işik, Muhittin und Hasan Dağ (2020). „The impact of text preprocessing on the prediction of review ratings“. In: *Turkish Journal of Electrical Engineering and Computer Sciences* 28.3, S. 1405–1421. DOI: https://doi.org/10.3906/elk-1907-46. Letzter Abruf am 19. Januar 2025 um 19:01 Uhr.

James, Gareth u. a. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer Nature. URL: https://hastie.su.domains/ISLP/ISLP_

website.pdf.download.html. Letzter Abruf am 19. Januar 2025 um 19:12 Uhr.

Maimon, Oded und Lior Rokach (2010). „Introduction to Knowledge Discovery and Data Mining“. In: *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, S. 1–15. ISBN: 978-0-387-09823-4. DOI: https://doi.org/10.1007/978-0-387-09823-4_1. Letzter Abruf am 19. Januar 2025 um 18:23 Uhr.

Naveed, Humza u. a. (2023). „A comprehensive overview of large language models“. In: *arXiv*. DOI: https://doi.org/10.48550/arXiv.2307.06435. Letzter Abruf am 19. Januar 2025 um 18:52 Uhr.

Raschka, Sebastian und Vahid Mirjalili (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikitlearn, and TensorFlow 2*. Packt publishing ltd.

Russell, Stuart und Peter Norvig (2012). *Künstliche Intelligenz: ein moderner Ansatz*. Bd. 3. Pearson, Higher Education (Always learning). ISBN: 978-3-86894-098-5.

Sanh, V (2019). „DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter“. In: *arXiv*. DOI: https://doi.org/10.48550/arXiv.1910.01108. Letzter Abruf am 19. Januar 2025 um 18:37 Uhr.

Schimanski, Tobias u. a. (2023). „ClimateBERT-NetZero: Detecting and Assessing Net Zero and Reduction Targets“. In: Swiss Finance Institute Research Paper No. 23-110. DOI: https://dx.doi.org/10.2139/ssrn.4599483. Letzter Abruf am 19. Januar 2025 um 18:02 Uhr.

Zhao, Huaqin u. a. (2024). „Revolutionizing Finance with LLMs: An Overview of Applications and Insights“. In: *arXiv*. DOI: https://doi.org/10.48550/arXiv.2401.11641. Letzter Abruf am 19. Januar 2025 um 19:33 Uhr.

Anhang

A.1 Übersicht der Kernbereiche großer Sprachmodelle 37
A.2 Trendentwicklung von LLMs im Kontext 38
A.3 Exemplarisches Klassifikationsbeispiel - Mensch und Modell 40
A.4 Grafische Gesamtübersicht aller Modellergebnisse 41
A.5 Gesamtübersicht der LLM Konfusionsmatrizen und Metriken 42
A.6 Klassifizierungsanleitung 43
A.7 Skripte 44

A.1 Übersicht der Kernbereiche großer Sprachmodelle

Systematische Übersicht der Kernbereiche großer Sprachmodelle. Hierbei teilt Abbildung A.1 die Bereiche Entwicklung und Anwendung in sieben Hauptbereiche auf: *Pre-Training*, *Fine-Tuning*, *Effizienz*, *Evaluation*, *Inference*, *Anwendungen* und *Herausforderungen*. Jeder dieser Bereiche fasst spezifische Konzepte oder Techniken zusammen, wobei der in Rot hervorgehobene Pfad, den Hauptbereich der Evaluation für diese Arbeit abbildet.

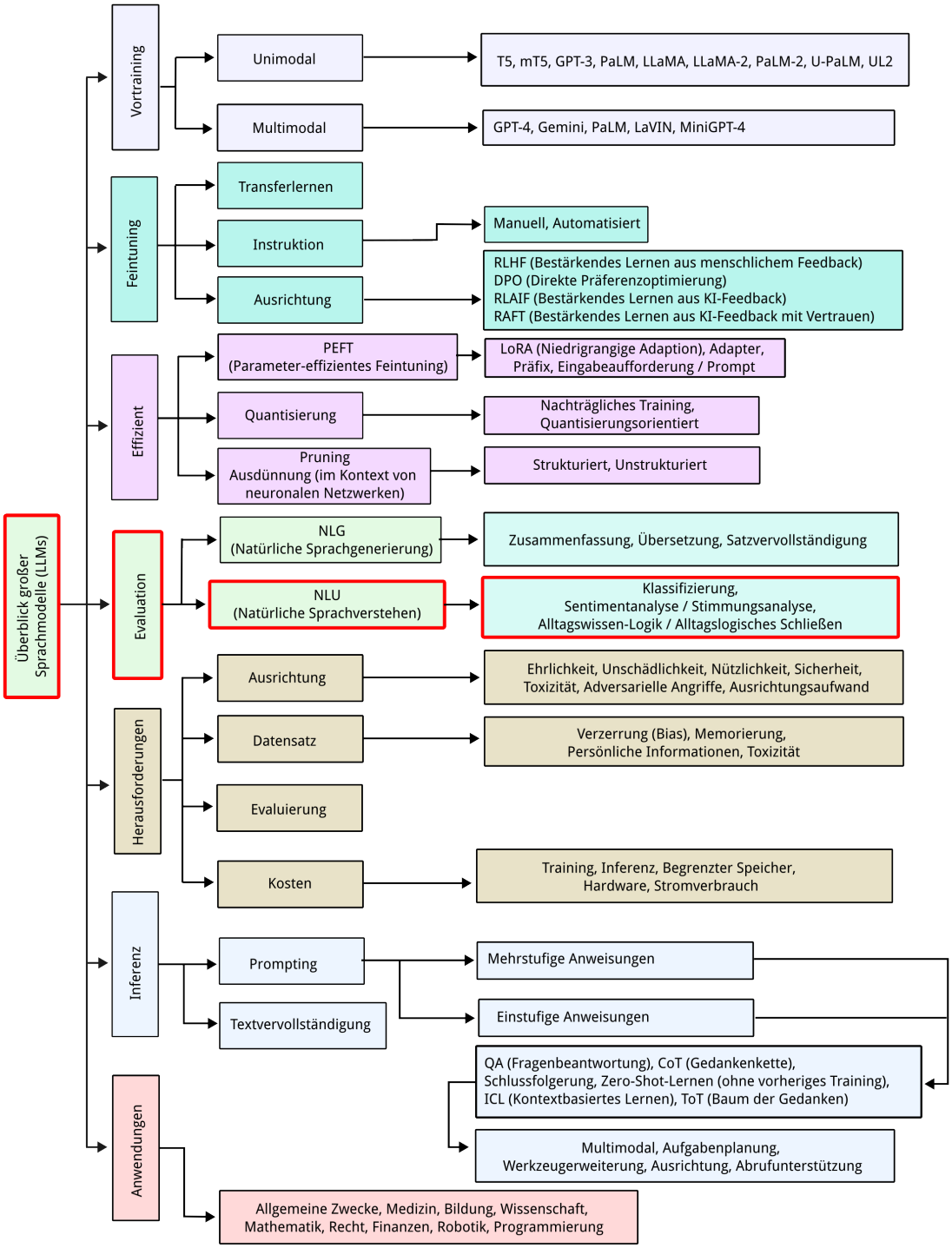

Abbildung A.1: Umfassende Übersicht von LLMs, die in sieben Bereiche unterteilt wurden: 1. Vortraining, 2. Feintuning, 3. Effizienz, 4. Inferenz, 5. Evaluation, 6. Anwendungen, 7. Herausforderungen.
Eigene Darstellung, in Anlehnung an: Naveed u. a. (2023).

A.2 Trendentwicklung von LLMs im Kontext

Abbildung A.2 bildet den Trend der über die Jahre veröffentlichten wissenschaftlichen Artikel im direktbezug zu den Schlüsselwörtern LLM, LLM+Fine-Tuning und LLM+Aligment.

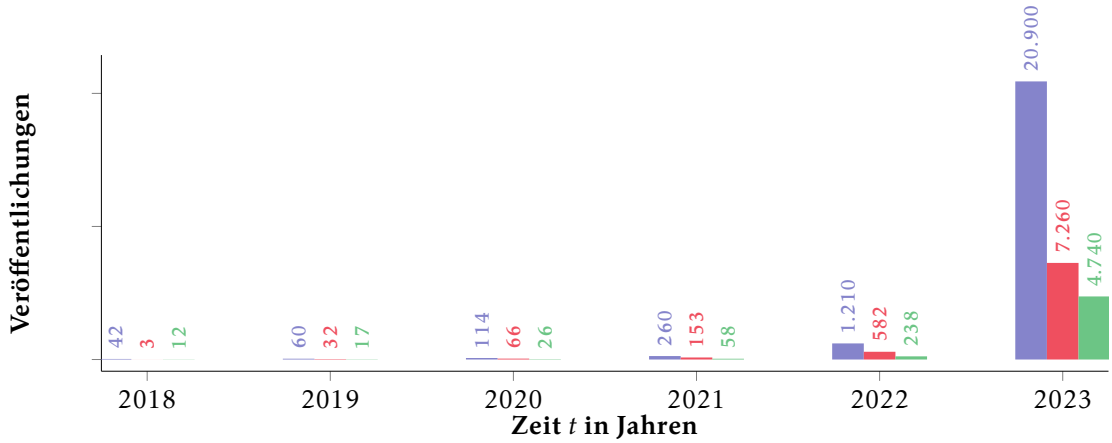

Abbildung A.2: Trend der über die Jahre veröffentlichten wissenschaftlichen Artikel im direktbezug zu den Schlüsselwörtern LLM, LLM+Fine-Tuning und LLM+Aligment. Hierbei bilden die blauen Balken die Summe der publizierten wissenschaftlichen Artikel mit dem Schlüsselwort **LLM** pro Jahr, die roten Balken **LLM+Fine-Tuning** und die grünen Balken **LLM+Alignment** pro Jahr.
Eigene Darstellung, in Anlehnung an: Naveed u. a. (2023); S.1

Die in Abbildung A.3 aufgetragenen blauen Flächen repräsentieren vortrainierte Modelle, während die orangefarbenen Flächen instruiert-trainierte Modelle kennzeichnen. Modelle in der oberen Hälfte stehen für *Open-Source*-Modelle, wohingegen die Modelle in der unteren Hälfte für *Closed-Source*-Modelle stehen. Die Abbildung verdeutlicht den zunehmenden Trend hin zu *instruiert-trainierte* Modellen und *Open-Source*-Modellen und hebt die sich entwickelnde Landschaft sowie die Trends in der Forschung zur natürlichen Sprachverarbeitung hervor. (vgl. Naveed u. a. (2023))

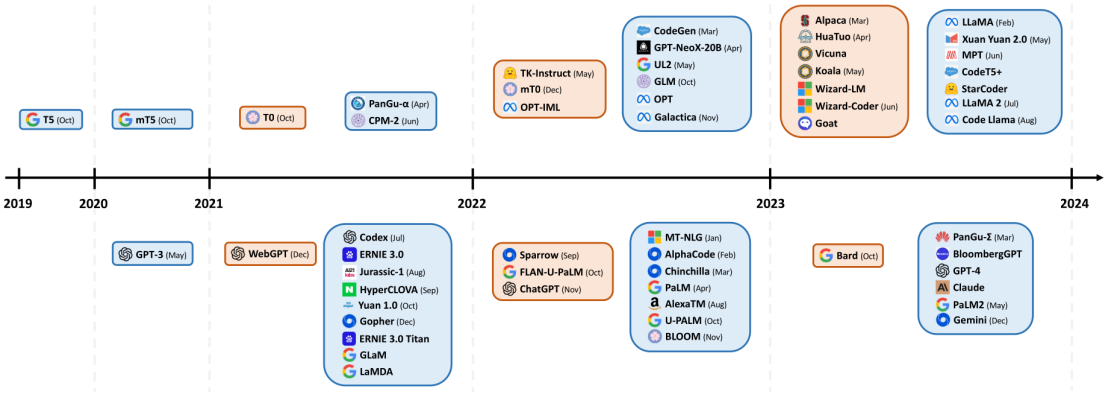

Abbildung A.3: Chronologische Darstellung der LLM-Veröffentlichungen.
Quelle: Naveed u. a. (2023); S.2.

A.3 Exemplarisches Klassifikationsbeispiel - Mensch und Modell

Abbildung A.4 zeigt einen exemplarischen Klassifikationsprozess, wobei in dem grauen Feld der Artikelauszug des Datensatzes steht, in dem grünen Feld der Klassifikationsprozess des *Llama 3.1*-Modells und in dem orangen Feld die Expertenklassifikation abgebildet ist. Hierbei wird exemplarisch aufgezeigt, wie verschieden Mensch und Modell trotz gleicher Aufgabenanleitung Bewerten können.

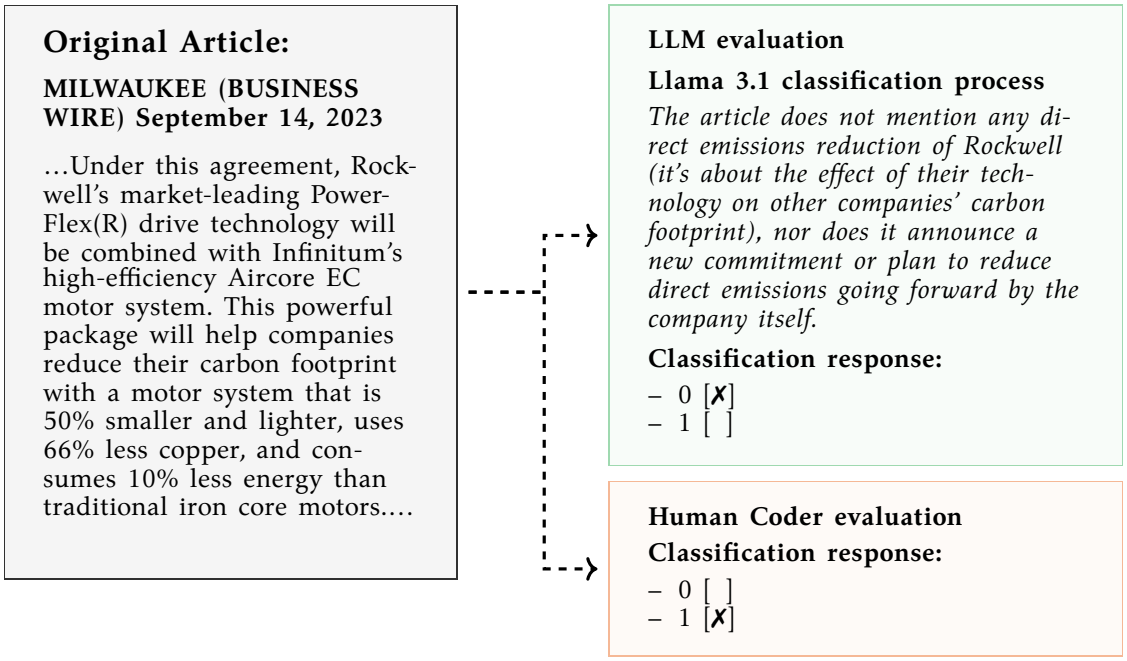

Abbildung A.4: Exemplarisches Klassifikationsbeispiel: LLM (Llama 3.1) & Experte am Beispiel des Original-Artikelauszugs (Artikel 982 von 1.000) des Datensatzes von Bauer u. a. (Nov 2024).

A.4 Grafische Gesamtübersicht aller Modellergebnisse

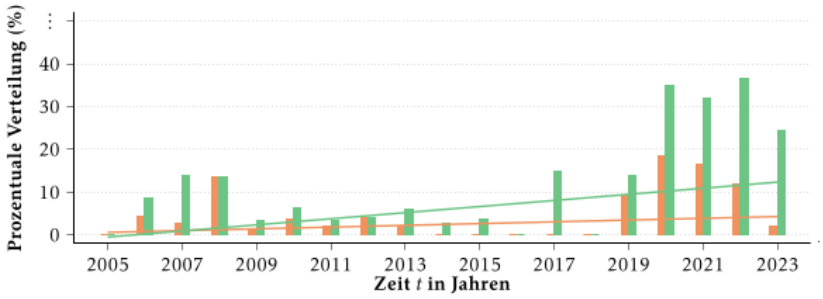

Abbildung 4.2: Prozentuale Verteilung positiver Artikel-Klassifikationen gemäß den Kategorien *Human-Coder* und *GPT-4*, die ein *neues* unternehmerisches Dekarbonisierungsversprechen von Treibhausgasemissionen repräsentieren.

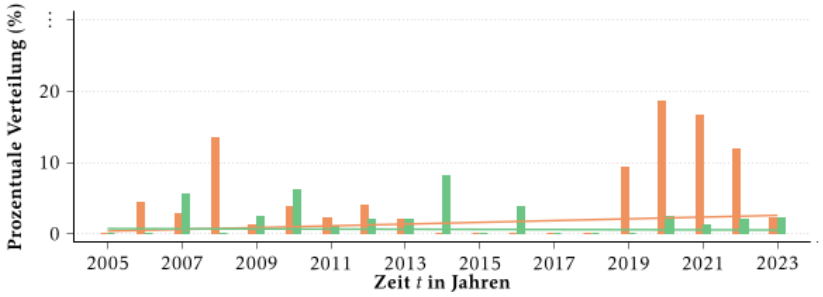

Abbildung 4.3: Prozentuale Verteilung positiver Artikel-Klassifikationen gemäß den Kategorien *Human-Coder* und *Gemma 2*, die ein *neues* unternehmerisches Dekarbonisierungsversprechen von Treibhausgasemissionen repräsentieren.

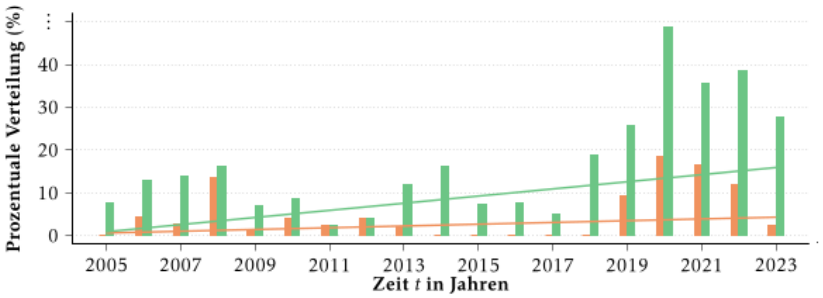

Abbildung 4.4: Prozentuale Verteilung positiver Artikel-Klassifikationen gemäß den Kategorien *Human-Coder* und *Llama 3.1*, die ein *neues* unternehmerisches Dekarbonisierungsversprechen von Treibhausgasemissionen repräsentieren.

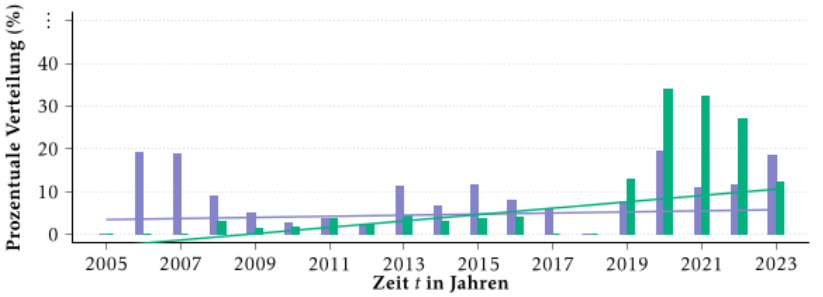

Abbildung 4.5: Prozentuale Verteilung positiver Artikel-Klassifikationen gemäß den Kategorien *Reduktion* und *Net-Zero*, die unternehmerische Dekarbonisierungsversprechen von Treibhausgasemissionen repräsentieren.

Abbildung A.5: Grafische Gesamtübersicht der Modellergebnisse.

A.5 Gesamtübersicht der LLM Konfusionsmatrizen und Metriken

Übersicht aller LLM-Konfusionsmatrizen mit zugehörigen Ergebnissmetriken.

		Human Coder		
		Negativ	**Positiv**	Total
GPT-4	**Negativ**	842	15	857
	Positiv	100	43	143
	Total	942	58	1.000

Precision:	0,30	**Recall:**	0,74
F_1 score:	0,43	**Accuracy:**	0,89

Abbildung A.6: Übersicht der GPT-4-Modellergebnisse in Form einer Konfusionsmatrix mit Metriken.

		Human Coder		
		Negativ	**Positiv**	Total
Gemma 2	**Negativ**	917	58	975
	Positiv	25	0	25
	Total	942	58	1.000

Precision:	0,00	**Recall:**	0,00
F_1 score:	0,00	**Accuracy:**	0,92

Abbildung A.7: Übersicht der Gemma-2-Modellergebnisse in Form einer Konfusionsmatrix mit Metriken.

		Human Coder		
		Negativ	**Positiv**	Total
Llama 3.1	**Negativ**	800	18	818
	Positiv	142	40	182
	Total	942	58	1.000

Precision:	0,22	**Recall:**	0,69
F_1 score:	0,33	**Accuracy:**	0,84

Abbildung A.8: Übersicht der Llama-3.1-Modellergebnisse in Form einer Konfusionsmatrix mit Metriken.

A.6 Klassifizierungsanleitung

Im Folgenden wird die Klassifizierungsanleitung vorgestellt, welche den Experten zur Verfügung gestellt wurde. Diese Anleitung soll dazu dienen, Erklärungen von freiwillige neuen Emissionsreduktionen in den zugeteilten US-amerikanischen Nachrichtenartikeln zu identifizieren. Hierzu haben Bauer u. a. (Nov 2024) folgende Anweisungen aufgestellt:

> In a new research project, we are investigating the financial and environmental effects of corporate announcements of decarbonization commitments. To identify these announcements from corporate news articles, we need your help! A decarbonization commitment is defined as follows: A firm makes a clear, action- able commitment to significantly reduce future greenhouse gas emissions. Green- house gases include carbon dioxide (CO_2) and methane.
>
> In the attached Excel list, you will find a random selection of news articles from Dow Jones (DJ newswires and Wall Street Journal articles). We are asking you to please identify decarbonization commitments in these articles. Please read carefully through each article, decide whether it constitutes a de- carbonization commitment. Then label it accordingly with "yes" (positive: the announcement contains a decarbonization commitment) or "no" (negative—no decarbonization commitment).
>
> Please classify an article as positive only if the company announces a signifi- cant reduction of direct emissions, that is, emissions that occur from sources controlled or owned by the company. The announcement should be news and should describe the company's commitments and plans for the future. Do not classify articles as positive that only contain announcements to reduce indirect emissions, that is, emissions that a company causes indirectly from the energy it purchases and uses. Also do not classify articles as positive if they are only about past performance, about a corporate social responsibility (CSR) report describing past emission reductions, about other environmental measures such as waste reduction, use of recycled paper, or planting trees, or announcements by the government. If an article is empty, or does not contain enough information, classify it as negative.

A.7 Skripte

Programmausdruck A.1: LLM Python setup code - Gemma 2.

```
#!/usr/bin/env python
from llama_index.llms.ollama import Ollama
from llama_index.core.llms import ChatMessage
from tqdm import tqdm
import pandas as pd

df = pd.read_excel('/PATH/scripte/human_sample.xlsx')
llm = Ollama(model="gemma2:latest", request_timeout=120.0)
responses = [] # Empty list for the LLM responses (classifications)

# Iterate through each row in the DataFrame
for index, row in tqdm(df.iterrows()):
  # Define the article content from the 'text' column
  article_text = row['text']
  # Prepare example messages prompt for the chat API, using the current article text[index]
  messages = [
    ChatMessage(
      role="system", content="""\
        Does the company in the following article meets all the criteria for a positive classification:
        * classify an article as positive only if The company announces a significant reduction of direct
        emissions, that is, emissions that occur from sources controlled or owned by the company. (most important)
        * It describes the company's commitments and plans for the future. (second most important)
        * It is a new commitment or plan to reduce direct emissions going forward, rather than just an
        announcement of past performance or other environmental measures. (third most important)
        * It does not contain announcements about indirect emissions or government-related activities.
        * It does not contain only about past performance, about a corporate social responsibility (CSR)
        report describing past emission reductions, about other environmental measures.

        Positive examples for a positive announcement would looks like the following four
        examples: Example 1: CenterPoint Energy introduces Carbon Policy committing to reductions
        in emissions. Goals build on 'companys ongoing commitment to reduce carbon emissions and
        use innovative technology to transition toward a cleaner energy future [. . . ] Leading
        U.S. energy delivery company CenterPoint Energy (NYSE: CNP) today announced a goal to
        reduce its operational emissions by 70 percent by 2035 and emissions attributable to
        natural gas usage in heating, appliances and equipment within the residential and
        commercial sectors by 20 to 30 percent by 2040. The 'companys reduction goals are based on
        its 2005 emissions.

        Example 2: Or another positive classification example would be like:
        July 21, 2020: Apple today unveiled its plan to become carbon neutral across its entire
        business, manufacturing supply chain, and product life cycle by 2030. The company is
        already carbon neutral today for its global corporate operations, and this new commit-
        ment means that by 2030, every Apple device sold will have net zero climate impact. [...]

        Example 3: October 14, 2009: Wells Fargo & Company (NYSE: WFC) announced today that it has
        set a goal to reduce its U.S-.based greenhouse gas emissions by 20 percent below 2008
        levels by 2018. The Company is focusing on reducing its carbon footprint as part of its
        continued environmental commitment to lead by example and to fulfill its pledge as a
        member of the U.S. Environmental Protection 'Agencys ('EPAs) Climate Leaders program,
        which Wells Fargo joined last year. [...]

        Example 4: January 16, 2020: Microsoft Corp. on Thursday announced an ambitious goal and a
        new plan to reduce and ultimately remove its carbon footprint. By 2030 Microsoft will be
        carbon negative, and by 2050 Microsoft will remove from the environment all the carbon the
        company has emitted either directly or by electrical consumption since it was founded
        in 1975. [...]

        Does the following article fit into a positive classification, answer only with 1, else 0:
      """
    ),
    ChatMessage(role="user", content=article_text),  # Pipe the article text
  ]
  # Send the request to the LLM
  resp2 = llm.chat(messages)
  # Extract the 'content' from the ChatResponse object
  response_content = resp2.message.content  # Access 'content' directly as an attribute
  response_string = str(response_content) # Convert response content to string and process it
  response_string = response_string.replace(".", "").lower() # Remove periods and convert to lowercase
  responses.append(response_string) # Append responses ('yes' or 'no') to the responses list

# Print 'responses' which contains all the 'yes' or 'no' classifications for each article
print(responses)
# Convert results to DataFrame and save to file
results_df = pd.DataFrame(responses)
results_df.to_csv("PATH/scripte/Gemma_2/Gemma_2_results.dat", index=False)
```

Programmausdruck A.2: LLM Python setup code - Llama 3.1.

```python
#!/usr/bin/env python
from llama_index.llms.ollama import Ollama
from llama_index.core.llms import ChatMessage
from tqdm import tqdm
import pandas as pd

df = pd.read_excel('/PATH/scripte/human_sample.xlsx')
llm = Ollama(model="llama3.1:latest", request_timeout=120.0)
responses = [] # Empty list for the LLM responses (classifications)

# Iterate through each row in the DataFrame
for index, row in tqdm(df.iterrows()):
  # Define the article content from the 'text' column
  article_text = row['text']
  # Prepare example messages prompt for the chat API, using the current article text[index]
  messages = [
    ChatMessage(
      role="system", content="""\
        Does the company in the following article meets all the criteria for a positive classification:
        * classify an article as positive only if The company announces a significant reduction of direct
        emissions, that is, emissions that occur from sources controlled or owned by the company. (most important)
        * It describes the company's commitments and plans for the future. (second most important)
        * It is a new commitment or plan to reduce direct emissions going forward, rather than just an
        announcement of past performance or other environmental measures. (third most important)
        * It does not contain announcements about indirect emissions or government-related activities.
        * It does not contain only about past performance, about a corporate social responsibility (CSR)
        report describing past emission reductions, about other environmental measures.

        Positive examples for a positive announcement would looks like the following four
        examples: Example 1: CenterPoint Energy introduces Carbon Policy committing to reductions
        in emissions. Goals build on 'companys ongoing commitment to reduce carbon emissions and
        use innovative technology to transition toward a cleaner energy future [. . . ] Leading
        U.S. energy delivery company CenterPoint Energy (NYSE: CNP) today announced a goal to
        reduce its operational emissions by 70 percent by 2035 and emissions attributable to
        natural gas usage in heating, appliances and equipment within the residential and
        commercial sectors by 20 to 30 percent by 2040. The 'companys reduction goals are based on
        its 2005 emissions.

        Example 2: Or another positive classification example would be like:
        July 21, 2020: Apple today unveiled its plan to become carbon neutral across its entire
        business, manufacturing supply chain, and product life cycle by 2030. The company is
        already carbon neutral today for its global corporate operations, and this new commit-
        ment means that by 2030, every Apple device sold will have net zero climate impact. [...]

        Example 3: October 14, 2009: Wells Fargo & Company (NYSE: WFC) announced today that it has
        set a goal to reduce its U.S-.based greenhouse gas emissions by 20 percent below 2008
        levels by 2018. The Company is focusing on reducing its carbon footprint as part of its
        continued environmental commitment to lead by example and to fulfill its pledge as a
        member of the U.S. Environmental Protection 'Agencys ('EPAs) Climate Leaders program,
        which Wells Fargo joined last year. [...]

        Example 4: January 16, 2020: Microsoft Corp. on Thursday announced an ambitious goal and a
        new plan to reduce and ultimately remove its carbon footprint. By 2030 Microsoft will be
        carbon negative, and by 2050 Microsoft will remove from the environment all the carbon the
        company has emitted either directly or by electrical consumption since it was founded
        in 1975. [...]

        Does the following article fit into a positive classification, answer only with 1, else 0:
      """
    ),
    ChatMessage(role="user", content=article_text),  # Pipe the article text
  ]
  # Send the request to the LLM
  resp2 = llm.chat(messages)
  # Extract the 'content' from the ChatResponse object
  response_content = resp2.message.content  # Access 'content' directly as an attribute
  response_string = str(response_content) # Convert response content to string and process it
  response_string = response_string.replace(".", "").lower() # Remove periods and convert to lowercase
  responses.append(response_string) # Append responses ('yes' or 'no') to the responses list

# Print 'responses' which contains all the 'yes' or 'no' classifications for each article
print(responses)
# Convert results to DataFrame and save to file
results_df = pd.DataFrame(responses)
results_df.to_csv("PATH/scripte/Llama_3.1/Llama_3.1_results.dat", index=False)
```

Programmausdruck A.3: LLM Python setup code - ClimateBERT-NetZero.

```python
#!/usr/bin/env python
from transformers import AutoModelForSequenceClassification, AutoTokenizer, pipeline
from transformers.pipelines.pt_utils import KeyDataset
import pandas as pd
import matplotlib.pyplot as plt
from tqdm.auto import tqdm

# Dataset and model names
dataset_name = "climatebert/climate_detection"
detector_model_name = "climatebert/distilroberta-base-climate-detector"
netzero_model_name = "climatebert/netzero-reduction"

# Load dataset
file_path = "PATH/scripte/human_sample.xlsx"
dataset = pd.read_excel(file_path)

# Ensure the timestamp column is in datetime format
dataset['timestamp'] = pd.to_datetime(dataset['timestamp'], errors='coerce')

# Extract the year from the timestamp column
dataset['year'] = dataset['timestamp'].dt.year

# Specify the column containing text data
text_column = "text"
timestamp_column = "timestamp"

# Step 1: Climate context classification (Climate Detector)
detector_model = AutoModelForSequenceClassification.from_pretrained(detector_model_name)
detector_tokenizer = AutoTokenizer.from_pretrained(detector_model_name, max_len=512)
detector_pipe = pipeline("text-classification", model=detector_model, tokenizer=detector_tokenizer, device=0)

## Filter for climate-relevant paragraphs
count = 0
climate_relevant_texts = []

for i, out in enumerate(tqdm(detector_pipe(dataset[text_column].to_list(), padding=True, truncation=True))):
  # Extract year for the current row
  year = dataset['year'].iloc[i]  # Use .iloc to get the year by index
  if out['label'] == "yes":  # Adjust the label if needed
    climate_relevant_texts.append((i, dataset[text_column][i], year))
    print(f"Article Nr.: {i+1}, Year: {year}, Classification: {out}")
  elif out['label'] == "no":
    count += 1

print("Number of classifications with label yes: ", len(dataset[text_column])-count)
print(f"Number of classifications with label no: {count}")

# Initialize a list to store results
results = []
if climate_relevant_texts:
  # Extract relevant indices, texts, and years
  relevant_indices, relevant_texts, relevant_years = zip(*climate_relevant_texts)

  # Load ClimateBERT-NetZero model
  netzero_model = AutoModelForSequenceClassification.from_pretrained(netzero_model_name)
  netzero_tokenizer = AutoTokenizer.from_pretrained(netzero_model_name, max_len=512)
  netzero_pipe = pipeline("text-classification", model=netzero_model, tokenizer=netzero_tokenizer, device=0)

  # Classify climate-relevant texts with NetZero model
  for idx, text, year, out in zip(relevant_indices, relevant_texts, relevant_years,
        tqdm(netzero_pipe(list(relevant_texts), padding=True, truncation=True))):
    results.append({"Index": idx + 1, "Year": year, "Prediction": out["label"]})
    print(f"Original Index: {idx+1}, Year: {year}, NetZero Prediction: {out}")
else:
  print("No climate-relevant paragraphs detected.")

# Convert results to DataFrame
results_df = pd.DataFrame(results)
#print(results_df)
results_df.to_csv("PATH/scripte/ClimateBERT_results_data.dat", index=False)
```