

Politische Texte zum Thema Klimawandel automatisch identifizieren: Ein XGBoost Modell

Memminger, Ronja

memminger@uni-potsdam.de
Universität Potsdam, Deutschland

Benndorf, Dietmar

dietmar.benndorf@uni-potsdam.de
Universität Potsdam, Deutschland

Stede, Manfred

stede@uni-potsdam.de
Universität Potsdam, Deutschland

Um die Diskurse zu einem gesellschaftlich relevanten Thema verfolgen zu können, müssen Einzeltexte daraufhin klassifiziert werden, ob sie das fragliche Thema adressieren oder nicht. Wir wählen hier die Thematik des Klimawandels und machen ein entsprechendes Klassifikationsmodell frei verfügbar.

Für das Englische existiert mit ClimateBERT (Webersinke et al., 2022) bereits ein solcher Klassifikator, während für die deutsche Sprache unseres Wissens bislang kein entsprechendes System bereitsteht. Unser Beitrag erläutert, wie wir unser Modell trainiert und getestet haben. Es ist ab sofort auf GitHub abrufbar: <https://github.com/rmemminger/klima-klassifikator>.

Hintergrund

Als Ausgangspunkt, also eine Quelle von deutschsprachigen "Klimatexten", wählen wir das "AfD Climate Change Corpus" (Stede und Memminger, 2025). Es besteht aus 174.270 tokens von Textmaterial, das von Funktionsträgern oder Organen der AfD publiziert wurde, u.a. Reden im Bundestag und im Europäischen Parlament, Pressemitteilungen und Beiträgen auf sozialen Medien. Texte wurden in Abschnitte unterteilt und diese dann anhand einer Stichwortsuche thematisch gefiltert. Die Stichworte waren:

klima, erwärmung, treihbaus, co2, kohle, energiewende, verkehrswende, fridays for future, extinction rebellion

Die Arbeit von Stede und Memminger (2025) zeigt beispielhaft einige automatische Analysen auf dem AfD-Klimawandel Korpus, namentlich die Merkmale populistischer Sprache und von verschiedenen positiven und negativen Emotionen, wobei die AfD Texte mit Textmaterial

dreier anderer Parteien (CDU/CSU, SPD), ebenfalls zum Thema Klimawandel, verglichen werden. (Diese stammen aus dem von Schaefer et al. (2023) erstellten Korpus.)

Aus diesem Korpus erstellten wir das Trainingsmaterial für unseren Klassifikator. 3200 Absätze wurden zufällig aus allen Datenquellen entnommen, wovon 3000 mindestens eines der o.g. Stichworte enthielten, und 200 frei von diesen Stichworten waren. Alle Texte wurden manuell daraufhin bewertet, ob sie dem Thema Klimawandel zuzuordnen sind. Unvollständige Texte oder solche, die nicht eindeutig zugeordnet werden konnten, wurden entfernt. Die resultierende Verteilung war 2374 relevante und 775 nicht relevante Texte.

Der Klassifikator besteht aus einem XGBoost Classifier (Chen und Guestrin, 2016). Er wird auf den annotierten Absätzen mit 200 estimators, einer maximalen Baumtiefe von 10, und einer Lernrate von 0.1 trainiert. Auf dem in-domain Test Set erzielt er einen F1-Score von 0,89.

Out-of-domain Test des Klassifikators

Das Anliegen unseres vorliegenden Beitrags ist, den Klassifikator als allgemeinen Text-Thema-Klassifikator zur Verfügung zu stellen. Dafür ist es erforderlich zu zeigen, dass das auf dem AfD-Korpus trainierte Modell nicht ausschließlich Klimatexte der AfD erkennt, sondern unabhängig der politischen Ausrichtung allgemein das Thema des Klimawandels erfasst. Dafür haben wir weiteres Textmaterial herangezogen, nämlich Bundestagsreden anderer Parteien im Umfang von 140 Texten (20 pro Partei inkl. fraktionslos). Im Rahmen der Evaluation wurde zunächst eine Erstannotation durchgeführt. Diese wurde anschließend vollständig in einer Zweitannotation überprüft. Die vollständige manuelle Annotation ergab eine Verteilung von 14 Klimatexten und 126 Nicht-Klimatexten. Diese Nicht-Balance entspricht dem Ziel dieser Phase, insbesondere die Precision des Klassifikators zu überprüfen.

Der Klassifikator erkannte 13 der 14 Klimatexte korrekt (1 False Negative) und klassifizierte einen Nicht-Klimatext fälschlich als Klima (1 False Positive). Daraus ergibt sich für die Klasse „Klima“ eine Precision von 0,93, ein Recall von 0,93 und ein F1-Score von 0,93. Über alle 140 Texte hinweg berechnet, liegt der Makro-F1-Score bei 0,96 und der Mikro-F1-Score bei 0,99.

Diese Ergebnisse zeigen, dass der Klassifikator auch auf parteiübergreifendem Material zuverlässig Klimatexte identifizieren kann.

Problematisierung und Ausblick

Texte zu politischen Themen sind unterschiedlich "einfach" mit Stichworten identifizierbar. Für ein Thema wie "Kohleausstieg" lässt sich recht einfach ein hoher Recall erzielen, weil ein Text kaum ohne Erwähnung von "Kohle",

"Ausstieg" oder "Kohleausstieg" auskommt; auch die Precision lässt sich durch Kombination der Stichworte gut optimieren (vgl. Müller-Hansen et al., 2021). Für "Klimawandel" ist ein guter Recall ebenfalls erreichbar und eine Stichwortsuche wurde für das Englische häufig praktiziert, z.B. von Hulme et al. (2018) für Kommentare aus 'Nature' und 'Science' - wo dann allerdings eine manuelle Durchsicht vorgenommen wurde, um viele false positives wieder zu eliminieren. Deshalb ist das Training eines automatischen Klassifikators für dieses Thema vorteilhaft: Sollen für eine neue Forschungsfrage und neue Korpora Klima-Texte identifiziert werden, ist das mit einem solchen Modell ohne nachträgliches manuelles Filtern möglich.

Dabei ist zu beachten, dass die thematischen Grenzen beim Klimawandel keineswegs eindeutig sind. Bei unseren manuellen Annotationen wurde u.a. klar, dass entschieden werden muss, ob Texte, die sich unterschiedlichen Aspekten der Energiewende widmen (und dabei möglicherweise "Klima" erwähnen, aber nicht zum Kernthema machen) einbezogen werden sollen oder nicht.

Unsere nächsten Schritte bestehen darin, den Klassifikator auf weiterem Textmaterial anderer Genres zu testen, insbesondere unterschiedlichen Arten von sozialen Medien (z.B. YouTube Untertitel und Kommentare), und sofern es sich als notwendig erweist, auf einem um solche Textsorten ergänzten, also wiederum manuell kuratierten, Korpus nachzutrainieren, um die Güte des Systems weiter zu verbessern.

Bibliographie

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 13, 785–94. <https://doi.org/10.1145/2939672.2939785>.

Hulme, Mike, Noam Obermeister, Samuel Randalls, and Maud Borie. 2018. "Framing the Challenge of Climate Change in Nature and Science Editorials". *Nature Climate Change* 8 (6): 515–21. <https://doi.org/10.1038/s41558-018-0174-1>.

Müller-Hansen, Finn, Max W. Callaghan, Yuan Ting Lee, Anna Leippbrand, Christian Flachsland, and Jan C. Minx. 2021. "Who Cares about Coal? Analyzing 70 Years of German Parliamentary Debates on Coal with Dynamic Topic Modeling". *Energy Research & Social Science* 72 (February): 101869. <https://doi.org/10.1016/j.erss.2020.101869>.

Schaefer, Robin, Christoph Abels, Stephan Lewandowsky, and Manfred Stede. 2023. "Communicating Climate Change: A Comparison Between Tweets and Speeches by German Members of Parliament". In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, edited by Jeremy Barnes, Orphée De Clercq, and Roman Klinger. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.wassa-1.42>.

Stede, Manfred, and Ronja Memminger. 2025. "AfD-CCC: Analyzing the Climate Change Discourse of a German Right-Wing Political Party". In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, edited by Katherine Atwell, Laura Biester, Angana Borah, et al. Association for Computational Linguistics. <https://aclanthology.org/2025.nlp4pi-1.14/>.

Webersinke, Nicolas, Mathias Kraus, Julia Anna Binger, and Markus Leippold. 2022. "ClimateBert: A Pretrained Language Model for Climate-Related Text". arXiv:2110.12010. Preprint, arXiv, December 17. <https://doi.org/10.48550/arXiv.2110.12010>.