Deep Learning-based Sentiment Analysis of Textual Data using ChatGPT Labeling

Exposé zur Bachelorarbeit

Orkan Soyyigit
Humboldt-Universität zu Berlin

7. Juli 2023

1 Einleitung

Mit zunehmender Vernetzung in unserer Welt gewinnt die Analyse von Textdaten eine immer größere Bedeutung. Diese Entwicklung ist größtenteils auf das kommerzielle Potenzial der Auswertung von massiven Datenmengen zurückzuführen, welche täglich auf Social Media generiert werden [1]. In diesem Kontext stellt die Sentimentanalyse - eine Disziplin des Natural Language Processings - ein wichtiges Werkzeug zur Untersuchung dieser enormen Datenmengen dar. Mit ihrer Hilfe können Stimmungen bzw. Haltungen in Texten erkannt werden, um Entscheidungen in Bereichen wie Marketing zu unterstützen [2].

Machine Learning ist in solchen Prozessen ein entscheidendes Verfahren, mit dessen Hilfe sich Muster und Zusammenhänge in diesen großen Datenmengen erkennen lassen können, um daraus letztendlich entsprechende Vorhersagen abzuleiten. Mindestens so wichtig ist aber auch die Qualität der Datensätze, mit denen solche Lernmethoden trainiert, getestet bzw. evaluiert werden. Hierbei ist zwischen Gold-Standard-Datensätzen und Silver-Standard-Datensätzen zu unterscheiden. Gold-Standard-Datensätze werden manuell von Experten annotiert und weisen eine hohe Qualität auf [3]. Diese genaue Annotation ist jedoch ein zeitaufwändiger und kostenintensiver Prozess, was die Verfügbarkeit und Größe der Datensätze limitiert [4]. Silver-Standard-Datensätze werden dagegen durch semiautomatisierte Verfahren erstellt bzw. annotiert und sind in der Regel deutlich größer [5]. Sie bieten damit eine kosteneffiziente Alternative - allerdings oft auf Kosten der Qualität.

Deep Learning - ein Teilbereich des Machine Learnings - hat in den vergangenen Jahren bedeutende Fortschritte gemacht und vielversprechende Ergebnisse in der Sentimentanalyse geliefert [6]. In einer weiterführenden Entwicklung hat sich insbesondere die Transformer-Architektur als sehr leistungsfähig erwiesen, welche erstmals in [7] vorgestellt wurde. In dieser Architektur kommen Mechanismen wie die sogenannte *Attention* zum Einsatz, um Zusammenhänge zwischen Wörtern in einem Text zu ermitteln, was zu einer verbesserten Leistung bei vielen NLP-Problemen führt [8].

Ein aktuelles Beispiel für die Anwendung der Transformer-Architektur ist das von OpenAI entwickelte ChatGPT, das speziell darauf ausgelegt ist, menschenähnlichen Text in einer dialogorientierten Form zu erzeugen [9]. Seine fortgeschrittenen Fähigkeiten zur Textgenerierung und Interaktion könnten es daher zu einem äußerst nützlichen Werkzeug für Aufgaben wie der automatisierten Annotation und der Analyse von Textdaten machen.

Diese Arbeit untersucht, wie gut sich ein kleiner Gold-Standard-Datensatz mit einem großen, durch ChatGPT gelabelten Silver-Standard-Datensatz anreichern lässt. Außerdem wird die Performance von Deep-Learning-Modellen für die Sentimentanalyse miteinander verglichen, wobei einerseits die Silver-Standard-Daten eines existierenden Korpus mit ChatGPT gelabelt werden und andererseits seine originalen Label zum Einsatz kommen. Anschließend wird die Leistung von ChatGPT bei der Labelzuweisung auf den Gold-Standard-Testdaten mit den Originaldaten sowie den zuvor erwähnten Modellen

verglichen.

2 Forschungsfragen

- Inwieweit verbessert die Ergänzung eines Gold-Standard-Datensatzes mit einem umfangreichen Silver-Standard-Datensatz die Leistung eines Sentimentanalyse-Modells?
- Wie beeinflusst die Verwendung von ChatGPT-generierten Label die Performance eines Deep-Learning-Modells für die Sentimentanalyse?
- Wie gut ist die Leistung von ChatGPT bei der Labelzuweisung auf den Gold-Standard-Testdaten im Vergleich zu den Originaldaten und spezifisch trainierten Deep-Learning-Modellen für die Sentimentanalyse?

3 Verwandte Arbeiten

Die Sentimentanalyse an sich ist ein weit erforschtes Gebiet, das im Kontext der maschinellen Textklassifikation und Datenannotation eine wichtige Rolle spielt. Mit der Einführung von ChatGPT im November 2022 hat dieser Bereich einen zusätzlichen Impuls erhalten.

Die folgenden Studien behandeln zwar nicht direkt die Fragestellungen dieser Arbeit, aber sie sind dennoch von hoher Relevanz, da sie die Fähigkeiten von ChatGPT bezüglich der Annotation und Erweiterung von Datensätzen untersuchen und bewerten.

Huang et al. haben die Leistung von ChatGPT bei der Annotation von 795 implizit hasserfüllten Tweets evaluiert [10]. Dabei hat ChatGPT jeweils als Antwort nicht nur die Label ausgegeben, sondern auch die Erklärungen, die seine Labelzuweisung rechtfertigen und begründen. Den Ergebnissen zufolge hat das Sprachmodell 80% der Tweets korrekt als hasserfüllt identifiziert. Die Untersuchung hat zudem gezeigt, dass die ursprünglichen Annotatoren ihre Bewertungen von Tweets signifikant änderten, als ihnen die jeweiligen Erklärungen des Sprachmodells vorgelegt wurden. Dies weist auf die potenzielle Einflussnahme der von ChatGPT erzeugten Erklärungen auf die Wahrnehmung und Entscheidungsfindung der menschlichen Annotatoren hin.

Gilardi et al. haben die Performance von ChatGPT mit menschlichen Codierern (Forschungsassistenten) und Crowdworkern (MTurk) in mehreren Annotationsaufgaben verglichen [11]. Bei der Analyse einer Stichprobe von 2382 Tweets, welche ursprünglich von den gleichen menschlichen Codierern erstellt wurde, hat sich gezeigt, dass ChatGPT in vier von fünf Aufgaben eine höhere Genauigkeit aufwies als die Crowdworker von MTurk. Zudem erzielte ChatGPT eine bessere Intercoder-Übereinstimmung als die besagten Annotatoren (Forschungsassistenten), die den Datensatz für diese Studie erneut annotiert haben - sprich ChatGPT ist konsistenter in seinen Ergebnissen und kommt häufiger zu den gleichen Klassifizierungen. Mit ihren Ergebnissen sind die Forscher zu dem Schluss gekommen, dass große Sprachmodelle wie ChatGPT das Potenzial haben, die Effizienz und Genauigkeit von Textklassifikationsaufgaben erheblich zu steigern.

Einen weiteren relevanten Bezugspunkt für diese Arbeit stellt die Studie von Zhong et al. dar [12]. Thematisch bestehen zwar einige Distanzen, da sich ihre Untersuchungen nicht speziell auf die Sentimentanalyse oder der Erweiterung von Datensätzen im Sinne von Gold-Standard und Silver-Standard mittels ChatGPT-Annotation konzentrieren - aber dennoch wird ein aufschlussreicher Vergleich zwischen ChatGPT und fein abgestimmten Modellen wie BERT [8] und RoBERTa [13] auf dem GLUE-Benchmark [14] durchgeführt. Dieser Benchmark umfasst verschiedene Bereiche wie beispielsweise Sentimentanalyse, Textähnlichkeit, Paraphrasierung und Natural Language Inference. Ihre Ergebnisse zeigen, dass ChatGPT in seiner Gesamtleistung mit dem BERT-Modell vergleichbar ist, aber vom RoBERTa-Modell übertroffen wird. Diese Befunde sind für die vorliegende Arbeit von großer Bedeutung, da hier ebenfalls vortrainierte Modelle wie BERT oder RoBERTa zum Einsatz kommen werden, um die nötigen Deep-Learning-Modelle zu erstellen und entsprechende Vergleiche bezüglich ihrer Performance zu ziehen.

4 Methodik

Um die in der Einleitung genannte Untersuchung anzustellen und die Forschungsfragen zu beantworten, wird ein mehrstufiger Ansatz verfolgt. Dieser Ansatz ist in Abbildung 1 dargestellt:

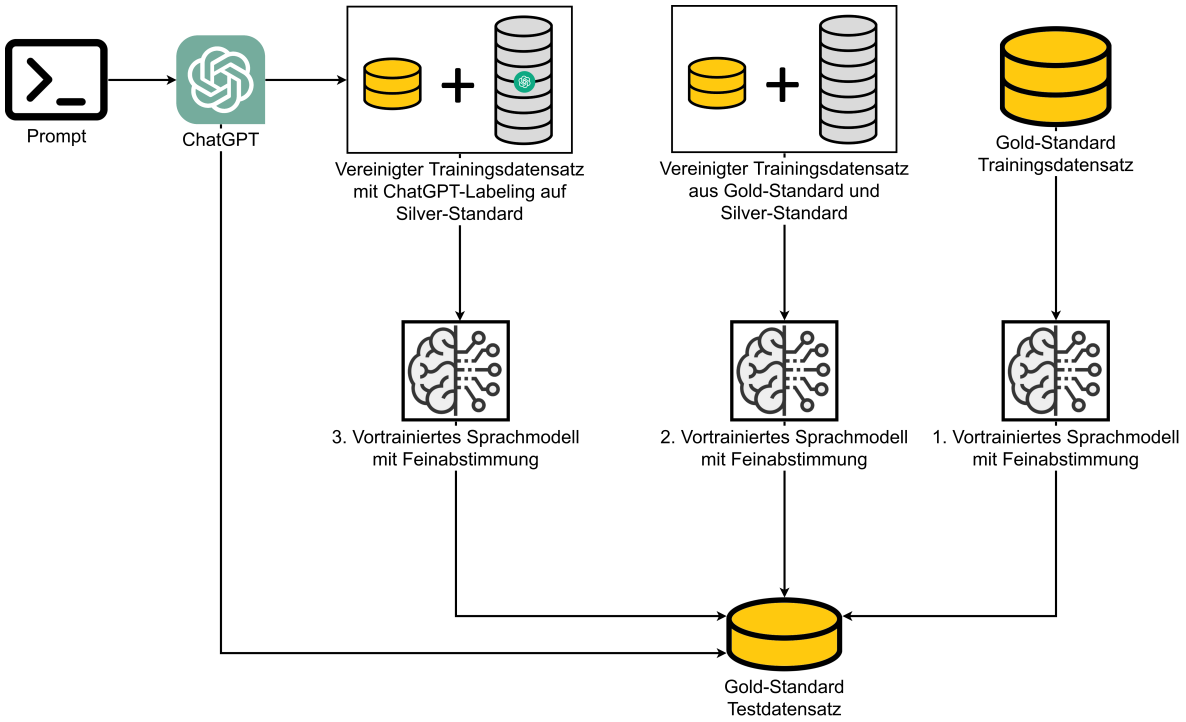

Abbildung 1: Zusammenstellung der Datensätze, Training der Modelle und Evaluation
Hinweis: Jeder Datensatz besteht immer aus zwei Spalten. Die erste Spalte enthält die Texteingabedaten und die zweite Spalte die entsprechenden Sentimentlabel [positiv,negativ], welche die Stimmung des jeweiligen Textes beschreiben.

Zu Beginn wird der Gold-Standard-Trainingsdatensatz verwendet, um damit das erste vortrainierte Sprachmodell zu trainieren bzw. fein abzustimmen. Anschließend wird dieses Modell auf dem Gold-Standard-Testdatensatz evaluiert, wobei die Leistung des Modells bezüglich Precision, Recall und F1-Score gemessen wird.

Parallel dazu wird ein vereinigter Datensatz erstellt, der sich aus dem Gold-Standard-Trainingsdatensatz und dem Silver-Standard-Trainingsdatensatz zusammensetzt. Mit diesem vereinigten Trainingsdatensatz wird das zweite vortrainierte Sprachmodell fein abgestimmt. Im Anschluss darauf wird das Sprachmodell ebenfalls auf dem Gold-Standard-Testdatensatz mit den entsprechenden Evaluationsmetriken bewertet.

Im nächsten Schritt wird eine spezielle Eingabeaufforderung bzw. Prompt definiert, welche an die ChatGPT-API gesendet wird. Diese Prompt sorgt dafür, dass alle Texteingabedaten des Silver-Standard-Trainingsdatensatzes an ChatGPT übergeben werden, wobei ChatGPT angeleitet wird, die Daten gemäß den vorgegebenen Sentimentklassen [positiv,negativ] zu labeln.
Die durch ChatGPT annotierten Silver-Standard-Trainingsdaten werden nun mit den Gold-Standard-Trainingsdaten vereinigt, womit das dritte vortrainierte Sprachmodell fein abgestimmt wird. Nach Abschluss des Trainings wird dieses Modell auf dem Gold-Standard-Testdatensatz evaluiert.

In einem zusätzlichen Schritt wird die gleiche Prompt an die ChatGPT-API gesendet - diesmal aber nicht mit den Silver-Standard-Trainingsdaten, sondern mit den Testdaten des Gold-Standards. Auf diese Weise wird das von ChatGPT generierte Labeling direkt mit den Gold-Standard-Label verglichen und bewertet.

Durch dieses Vorgehen kann die Leistung jedes Modells sowohl isoliert als auch im Vergleich zu den anderen Modellen und insbesondere zu den von ChatGPT erzeugten Label bewertet werden.

Literatur

[1] Jürgen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, Daniel M. Romero, Jahna Otterbacher, Carsten Schwemmer, Kenneth Joseph, David Garcia, and Fred Morstatter. Just another day on twitter: A complete 24 hours of twitter data. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1073–1081, Jun. 2023. URL https://ojs.aaai.org/index.php/ICWSM/article/view/22215.

[2] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023. ISSN 0167-8116. URL https://www.sciencedirect.com/science/article/pii/S0167811622000477.

[3] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. *CoRR*, abs/1912.00741, 2019. URL http://arxiv.org/abs/1912.00741.

[4] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? GPT-3 can help. *CoRR*, abs/2108.13487, 2021. URL https://arxiv.org/abs/2108.13487.

[5] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, 150, 01 2009. URL https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.

[6] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253, 2018. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1253.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[9] Debby R. E. Cotton, Peter A. Cotton, and J. Reuben Shipway. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, 0 (0):1–12, 2023. URL https://doi.org/10.1080/14703297.2023.2190148.

[10] Fan Huang, Haewoon Kwak, and Jisun An. Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*. ACM, apr 2023. URL https://doi.org/10.1145%2F3543873.3587368.

[11] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks, 2023. URL https://arxiv.org/abs/2303.15056.

[12] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert, 2023. URL https://arxiv.org/abs/2302.10198.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

[14] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461.