Humboldt-Universität zu Berlin

Institut für Informatik

Automatische Klassifikation von Alltagsszenarien mit einem BERT-basierten Siamese-Netzwerk

Exposé

von Tim Patzak

2. Juli 2021

1 Einleitung

Die automatische Analyse von Bewertungen, Stimmungen, Meinungen, Einstellungen und Emotionen in Texten sind viel untersuchte Felder in den letzten 10 Jahren. Sie werden unter den Begriffen Sentiment Analysis und Opinion Mining [7, 8] zusammengefasst. Der Hauptfokus lag dabei vor allem auf Texten aus den Bereichen Produktbewertungen [3, 9], Filmbewertungen [4, 10] oder politischen Meinungsäußerungen [5, 11]. Trotz der Umgangssprache und Abkürzungen in nutzergenerierten Inhalten aus populären Internetforen, wie *reddit*, *Quora* oder *tumblr*, lassen sich Informationen durch Opinion Mining und Sentiment Analysis systematisch erfassen.
Eine solche Informationsquelle ist der Subreddit *r/AmItheAsshole*[1]. Die Nutzer*innen schildern hier hypothetische oder reale Anekdoten mit ihren Interaktionen mit ihren Mitmenschen. Sie stellen den anderen Nutzern*innen des Forums die Frage, ob sie sich moralisch richtig oder falsch verhalten haben. In Form von Kommentaren können Nutzer*innen ihre Meinungen über die moralische Bewertung der Handlungen in der Anekdote abgeben. Sie wählen, ob die/der Autor*in, die andere Person, alle oder gar keiner sich nach ihren eigenen Normen falsch verhalten hat. Per Mehrheit wird entschieden wie der Beitrag auf dem Subreddit bewertet wird. Im Beispiel der Abbildung 1 tritt der Autor mit seinem Vater wieder in Verbindung, der 9 Jahre im Gefängnis saß. Er stellt ihn seiner Familie vor, die aber nichts mit dem Vater des Autors zu tun haben wollen. Die Kommentare mit den meisten Upvotes stimmen dafür, dass der Autor sich in dieser Situation falsch verhalten hat (siehe Abbildung 2), wodurch der Beitrag das Label „**Asshole**“ erhält. Der Kommentar gibt außerdem Anlass zur Vermutung, dass es sich bei dem Vater um einen Sexualstraftäter handelt.
Bisherige Arbeiten wie SCRUPLES [1] und die Bachelorarbeit von Scott Fletcher [19] befassen sich bereits mit diesem Thema. Mithilfe eines Korpus aus Anekdoten dieses Subbreddits trainieren sie einen Klassifikator, der das Label und damit die moralische Bewertung einer Anekdote versucht vorherzusagen. Lourie et al. [1] stellen den Korpus SCRUPLES frei zur Verfügung und präsentieren einige Baseline-Klassifikatoren. Beide Arbeiten konzentrieren sich ausschließlich auf die Anekdoten und berücksichtigen nicht die Kommentare. In dieser Arbeit soll untersucht werden, ob und inwiefern das Einbeziehen der Kommentare die Klassifikation solcher Situation verbessern kann. Die Vermutung liegt nahe, dass zusätzliche Kontextinformationen und die in den Kommentaren enthaltenen Argumente, wie aus Abbildung 2, die Klassifikation verbessert. Wie auch in SCRUPLES und der Arbeit von Scott Fletcher, soll ein Klassifikator basierend auf BERT [2] entwickelt und evaluiert werden.

[1] https://www.reddit.com/r/AmItheAsshole/

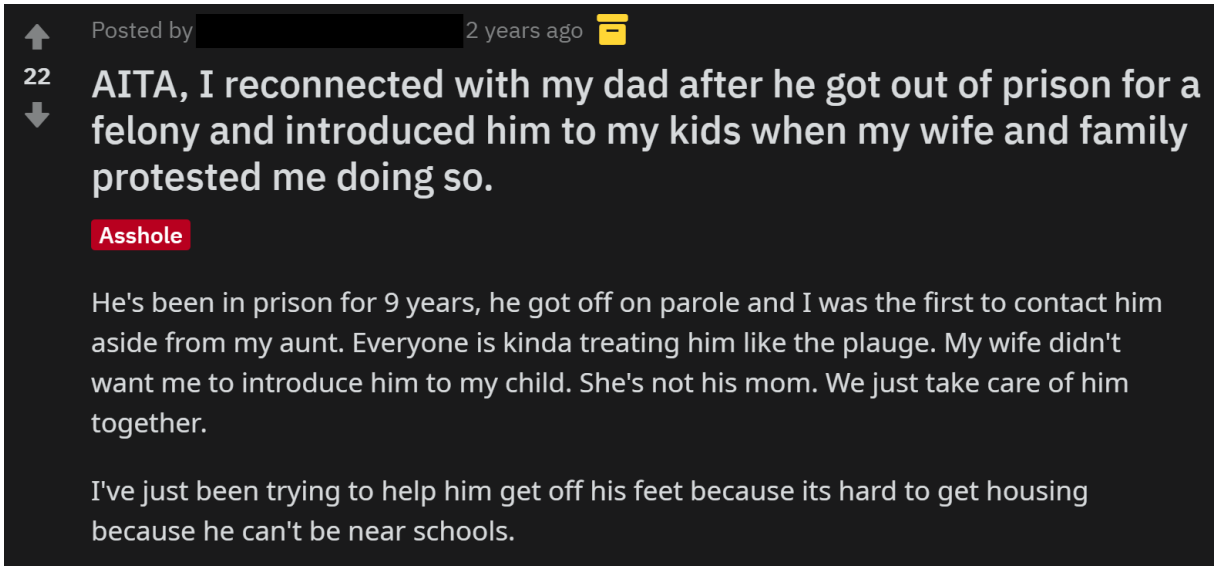

Abbildung 1: Beispiel einer Anekdote aus *r/AmItheAsshole* mit Label YTA (You're The Asshole) [32]

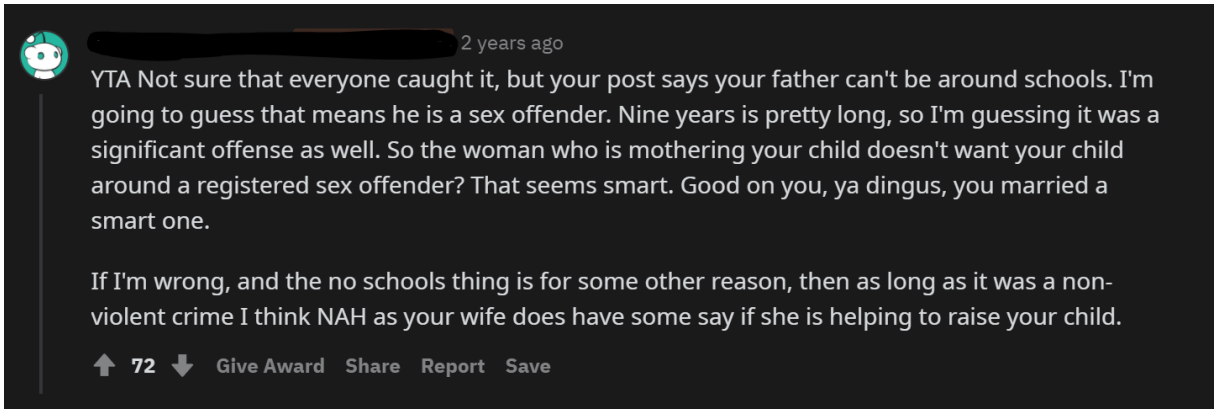

Abbildung 2: Kommentar über Anekdote aus Abbildung 1 mit negativer Bewertung

2 Verwandte Arbeiten

Reddit diente bereits als Plattform verschiedener wissenschaftlicher Arbeiten. Low et al. [12] verwenden verschiedene NLP-Techniken auf Beiträgen in Subreddits, die sich mit der geistigen Gesundheit beschäftigen. Sie stellen eine steigende Anzahl von Beiträgen bezüglich Stress, Einsamkeit und Suizidalität während der COVID-19 Pandemie fest. Vorallem bei Tabuthemen und Verhaltensweisen kann Reddit besonders nützlich sein. Greaves und Dykeman [13] verwenden Beiträge von Reddit, um sprach-

liche Hinweise zu finden, die auf Selbstverletzungen deuten.
In seiner Bachelorarbeit [19] erstellt Scott Fletcher einen Korpus von rund 75.000 Anekdoten aus dem Forum *r/AmITheAsshole*. Dabei verwendet er nicht die von Reddit erstellten Label für die Anekdoten, sondern aus den Kommentaren konstruierte, gewichtete Label - *positiv* und *negativ*. Auf diesem Korpus trainiert er eine Support Vector Machine (SVM) und einen auf BERT basierenden Klassifikator. Scott Fletcher erreicht dabei mit dem SVM-Ansatz eine Precision von 0,616 und einen Recall 0,617. Der BERT-basierte Klassifikator erreicht höhere Werte von 0,650 bzw. 0,649.
Neben dem Datensatz SCRUPLES, bestehend aus 32.000 Anekdoten aus *r/AmITheAsshole*, stellen Lourie et al. die Performance einiger Baseline Modelle auf den Daten von SCRUPLES vor. Darunter sind auch BERT [2] und RoBERTa [6]. Die Anekdoten in SCRUPLES beinhalten Metadaten wie *author*, *text*, *scores* und *label*. Im Gegensatz zu Reddit verwenden Lourie et al. leicht abgeänderte Label, wie in Tabelle 1 zu sehen ist. Mit RoBERTa [14] erreichen sie einen Macro-F1-Score von 0,305 als Baseline.

Reddit	**SCRUPLES**
YTA	AUTHOR
NTA	OTHER
ESH	EVERYBODY
NOH	NOBODY
INFO	INFO

Tabelle 1: Aufteilung der Anekdoten in SCRUPLES

Textklassifikation ist eine verbreitete Aufgabe im Bereich des Natural Language Processings mit dem Ziel, Texte in eine oder mehrere Kategorien einzuteilen. Typische Anwendungsbereiche sind Sentiment-Analyse in sozialen Medien [20, 21], Spamfilter [22, 23] und Kategorisierung von Nachrichtenartikeln [24, 25]. Zu den traditionellen Methoden der Textklassifikation gehören der Naive Bayes-Algorithmus [26, 27] und Support-Vektor-Maschinen (SVM) [28, 29]. In letzter Zeit haben Deep-Learning-Modelle traditionelle Modelle in vielen Textklassifizierungsaufgaben überholt [30]. Ein solches Modell ist BERT [2]. BERT verwendet das *Transfer-Learning*-Prinzip. Dabei werden Modelle auf großen Datensätzen ohne Labels vortrainiert (Pretraining) und anschließend kann eine Anpassung auf komplexere Aufgaben vorgenommen werden (Finetuning).

3 Konzeption und Vorgehen

3.1 Fragestellung

Bisherige Arbeiten beziehen sich in erster Linie auf die Anekdoten. In dieser Arbeit sollen zusätzlich die Kommentare einbezogen werden, um eine präzisere Klassifikation zu erreichen. In den Kommentaren auf *r/AmItheAsshole* sind nicht nur die Bewertungen des Szenarios zu finden, sie begründen in der Regel noch diese Bewertung. Die Nutzer liefern Argumente für die eine oder andere Richtung. Daher lässt sich vermuten, dass in den Kommentaren weitere Kontextinformationen und Hinweise zu finden sind, die nicht aus dem eigentlichen Beitrag hervorgehen. Verbessert also das Einbeziehen einer bestimmten Anzahl von Top-Kommentaren (Kommentare mit den meisten Upvotes) die Klassifikation der Anekdote?

3.2 Daten

Lourie et al. [1] stellen mit SCRUPLES einen Datensatz bestehend aus 32.000 Anekdoten in Form von nutzergenerierten Beiträgen aus dem Subreddit *r/AmITheAsshole* zur Verfügung. Jede Anekdote besteht aus drei Kernkomponenten: einem Titel *title*, einer Beschreibung *text* und den Bewertungen *scores*. Bewertungen unterteilen sich in die bereits beschriebenen Klassen **AUTHOR, OTHER, EVERYBODY, NOBODY** und **INFO**. Die Verteilung der Bewertungen ist in Tabelle 2 zu sehen. Die am häufigsten vorkommende Bewertung stellt gleichzeitig auch die Beschriftung des Beitrags *label* dar (siehe Abbildung 3). Hinzu kommen weitere Informationen, wie *id*, *type*, *action*, *type* und *binarized label*, die hier der Einfachheit halber im Beispiel der Abbildung 3 weggelassen worden sind.

AUTHOR	OTHER	EVERYBODY	NOBODY	INFO
29.8%	54.4%	4.8%	8.9%	2.1%

Tabelle 2: *label*-Verteilung der Anekdoten aus SCRUPLES

Die 32.000 Anekdoten des SCRUPLES Korpus sind bereits in Trainings-, Validierungs- und Testdaten eingeteilt. Die Kommentare der Anekdoten, die für diese Arbeit relevant sind, bietet SCRUPLES jedoch nicht. Mithilfe der ID der Beiträge und der Reddit-API **Pushshift** [31] können die Kommentare nachträglich gesammelt und den Beiträgen zugeordnet werden. Im Rahmen der Arbeit wird hierfür ein Tool zur Erweiterung des Korpus entwickelt.

3.3 Modell

Im Rahmen der Arbeit werden wir bestehende BERT-Modelle für die Klassifikation der Ankedoten einsetzen. Besondere Herausforderungen sind dabei die Eingabe mehrerer und längerer Texte, da BERT

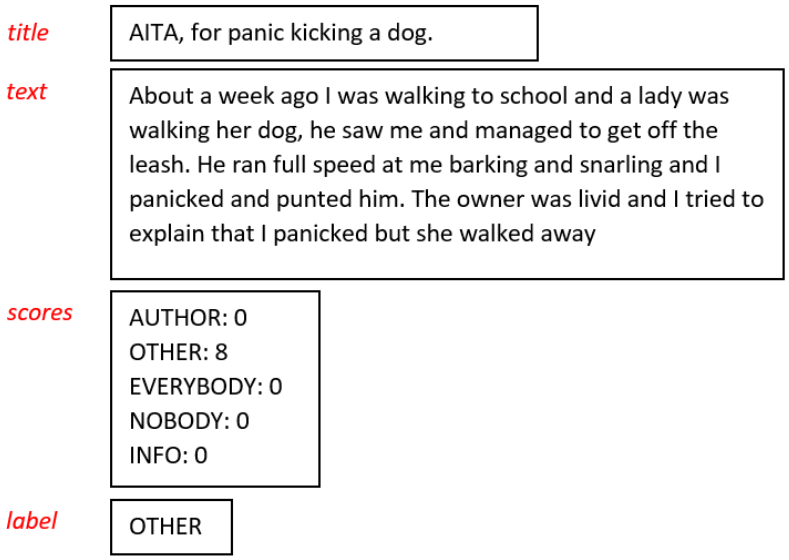

Abbildung 3: Beispiel einer Anekdote aus SCRUPLES mit dem *label* **OTHER**

maximal Sequenzen mit einer Länge von 512 Tokens unterstützt. Es gibt verschiedene Varianten mehrere Texte in BERT einzugeben [16–18]. Wir folgen hier dem Ansatz von Reimers und Gurevych [15], die mit Sentence BERT (SBERT) eine Siamese Netzwerkarchitektur vorschlagen. Diesem Ansatz nach werden die einzelnen Anekdoten oder Kommentare zunächst getrennt mit BERT-Netzwerken repräsentiert. Es entstehen Text Embeddings mit fester Größe für Texte von verschiedener Länge. Finetuning von BERT für die Klassifikation haben Reimers und Gurevych mithilfe der Softmax-Funktion vorgenommen. Die Repräsentation werden dann einem zusätzlichen Klassifikator in Form eines Feedforward Neural Networks übergeben. In Abbildung 5 ist die Architektur schematisch dargestellt.

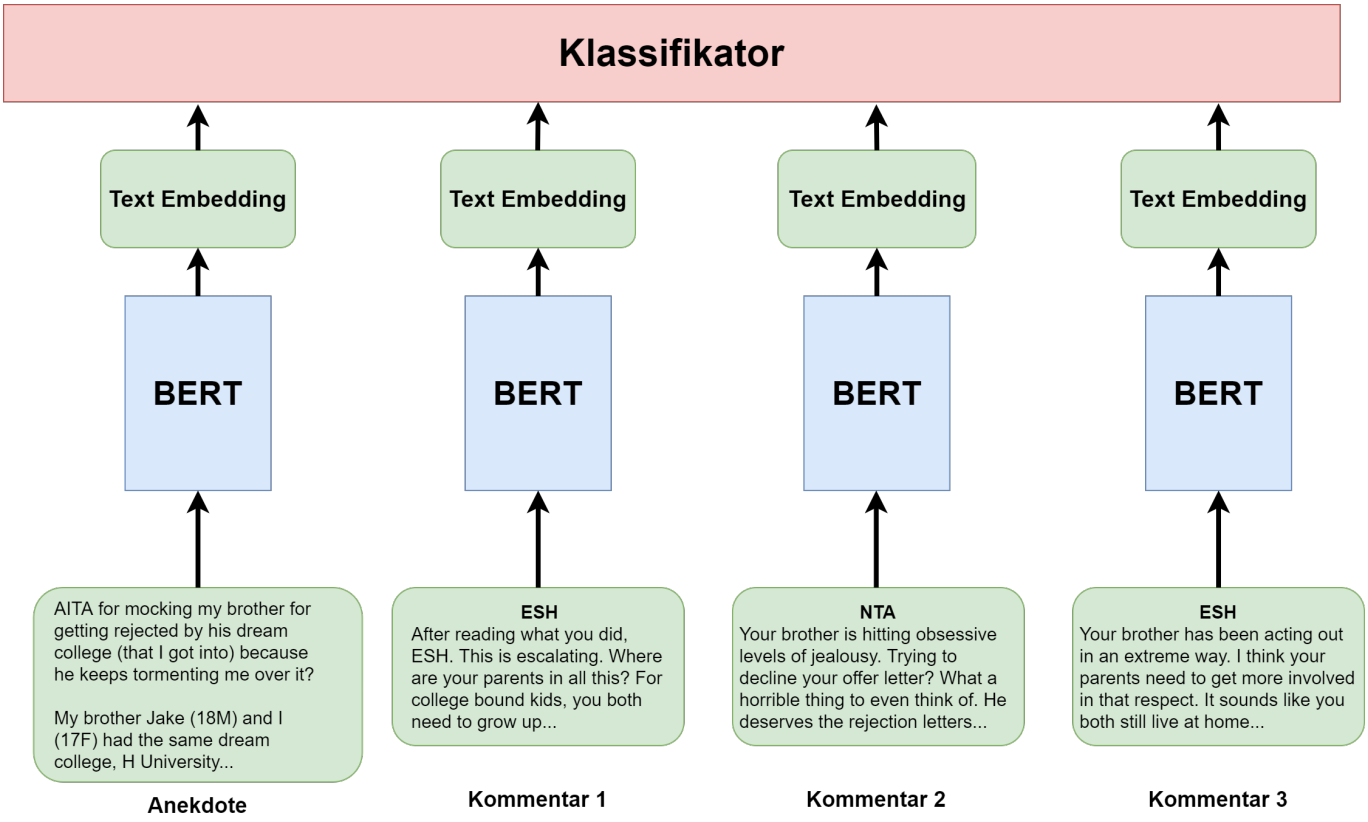

Abbildung 4: Übersicht des Modells mit Beispielen [33]
von Anekdote und Top-3 Kommentaren als gemeinsame Eingabe

4 Evaluation

Zusätzlich zu dem erweiterten Modell mit Kommentaren sollen zunächst die Ergebnisse aus Scott Fletchers Bachelorarbeit mit SCRUPLES als Datensatz reproduziert werden. Wir erstellen einen Klassifikator als Baseline, der nur den Text der Anekdote als Input erhält. Die Ergebnisse daraus können wir dann mit dem Klassifikator mit Kommentaren vergleichen. Zum Trainieren, Validieren und Evaluieren des Klassifikators, der die Kommentare als zusätzliche Features verwendet, soll der Datensatz von SCRUPLES verwendet werden. In Tabelle 3 sind die Splits der Daten, wie sie in SCRUPLES vorgenommen wurden, zu sehen. Es werden verschiedene Möglichkeiten der Integration der Kommentare untersucht. Eine Variante wäre die Top-n Kommentare mit den meisten Zustimmungen hinzuzuziehen, unabhängig davon welches *label* diese haben. Alternativ ließe sich für jedes *label* der Kommentar mit den meisten Zustimmungen bestimmen und hinzuziehen. Diese Varianten sollen auch untereinander verglichen werden. Als Maße zur Auswertung und Vergleich werden der F1-Score, Precision und Re-

call verwendet. Zudem soll eine qualitative Analyse der Daten vorgenommen werden. Dabei sollen die Unterschiede in den Vorhersagen der Modelle genauer untersucht werden.

TRAINING	VALIDIERUNG	TEST
27.766	2.500	2.500

Tabelle 3: Aufteilung der Anekdoten in SCRUPLES

Literatur

[1] Nicholas Lourie, Ronan Le Bras, Yejin Choi. Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes. arXiv:2008.09094. 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2018.

[3] K. L. Santhosh Kumar, Jayanti Desai, Jharna Majumdar. Opinion mining and sentiment analysis on online customer review. 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). 2016.

[4] Malini R., Sunitha M.R. Opinion Mining on Movie Reviews. 2019 1st International Conference on Advances in Information Technology (ICAIT). 2019.

[5] Mohd Zeeshan Ansari, M.B. Aziz, M.O. Siddiqui, H. Mehra, K.P. Singh. Analysis of Political Sentiment Orientations on Twitter. Procedia Computer Science, Volume 167. 2020.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. 2019.

[7] Shiliang Sun, Chen Luo, Junyu Chen. A review of natural language processing techniques for opinion mining systems. Information Fusion, Volume 36. 2017.

[8] Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, Gurpreet Kaur. Opinion mining and sentiment analysis. 3rd International Conference on Computing for Sustainable Global Development (INDIACom). 2016.

[9] Mohamad Syahrul Mubarok, Adiwijaya, and Muhammad Dwi Aldhi. Aspect-based sentiment analysis to review products using Naïve Bayes. AIP Conference Proceedings 1867, 020060. 2017.

[10] Atiqur Rahman, Md. Sharif Hossen. Sentiment Analysis on Movie Review Data Using Machine Learning Approach. International Conference on Bangla Speech and Language Processing (ICBSLP). 2019.

[11] Pawel Sobkowicza, Michael Kascheskya, Guillaume Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. Government Information Quarterly, Volume 29, Issue 4. 2012.

[12] Daniel M. Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, Satrajit S. Ghosh. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. Journal of Medical Internet Research 2020;22(10):e22635. 2020.

[13] Mandy M Greaves and Cass Dykeman. “A Corpus Linguistic Analysis of Public Reddit Blog Posts on Non-Suicidal Self-Injury”. In: arXiv preprint arXiv:1902.06689. 2019.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, volume: abs/1907.11692. 2019.

[15] Nils Reimers, Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084. 2019.

[16] Lee, Seolhwa, Joao Sedoc. Using the Poly-encoder for a COVID-19 Question Answering System. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. 2020

[17] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, Jason Weston. Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. arXiv:1905.01969. 2019.

[18] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, Ivan Vulić. Efficient Intent Detection with Dual Sentence Encoders. arXiv:2003.04807. 2020.

[19] Scott Fletcher. Assessing the Effectiveness of Text Classification Models for Moral Judgments using Reddit’s r/AmITheAsshole as a Data Set. Bachelor-Thesis. Humboldt-Universität zu Berlin. 2019.

[20] Alexandra Balahur. Sentiment Analysis in Social Media Texts. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Seiten: 120-128. 2013.

[21] C. Hutto, Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1). 2014.

[22] A. Harisinghaney, A. Dixit, S. Gupta and A. Arora, Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm,”2014 International Conference on Reliability Optimization and Information Technology (ICROIT), 2014, Seiten: 153-155. 2014.

[23] J. Clark, I. Koprinska and J. Poon, Ä neural network based approach to automated e-mail classification,Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003), 2003, Seiten: 702-705. 2003.

[24] Antonellis I., Bouras C., Poulopoulos V. (2006) Personalized News Categorization Through Scalable Text Classification. In: Zhou X., Li J., Shen H.T., Kitsuregawa M., Zhang Y. (eds) Frontiers

of WWW Research and Development - APWeb 2006. APWeb 2006. Lecture Notes in Computer Science, vol 3841. 2006.

[25] Conroy, N.K., Rubin, V.L. and Chen, Y. (2015), Automatic deception detection: Methods for finding fake news. Proc. Assoc. Info. Sci. Tech., 52: 1-4. 2015.

[26] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim and Sung Hyon Myaeng, SSome Effective Techniques for Naive Bayes Text Classification,ïn IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, Seiten: 1457-1466. 2006.

[27] H. Zhang and D. Li, Naïve Bayes Text Classifier,"2007 IEEE International Conference on Granular Computing (GRC 2007), 2007, Seiten: 708-708. 2007.

[28] S. Tong, D. Koller. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2(1), Seiten: 45–66. 2002.

[29] Wen Zhang, Taketoshi Yoshida, Xijin Tang. Text classification based on multi-word with support vector machine. Knowledge-Based Systems, Volume 21, Issue 8, 2008, Seiten: 879-886. 2008.

[30] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad Meysam Chenaghlu, Jianfeng Gao. Deep Learning Based Text Classification: A Comprehensive Review: arXiv:2004.03705. 2020.

[31] Webseite der Reddit-API Pushshift. `https://pushshift.io/`. letzter Zugriff: 28.05.2021 .

[32] Beispiel 1 eines Beitrags auf *r/AmItheAsshole*. `https://www.reddit.com/aphyew`. letzter Zugriff: 28.05.2021 .

[33] Beispiel 2 eines Beitrags auf *r/AmItheAsshole*. `https://www.reddit.com/ijcb69`. letzter Zugriff: 28.05.2021 .