netidee

STIPENDIEN

Stipendium- Jaqueline Böck

Zwischenbericht | Call 17 | Stipendium ID 6300

Inhalt

1 Einleitung

Bedrohungen, Beleidigungen, Hetze und hasserfüllte Beiträge sind einige der größten Herausforderungen, die durch die Digitalisierung entstehen. Derzeit werden viele soziale Plattformen wie Twitter, TikTok und Facebook von Menschen moderiert. Dieses moderieren ist aber oft sehr Zeit und Ressourcen aufwendig. Da auch die Datenmenge in der heutigen Zeit immer schneller und immer mehr zunimmt werden automatische Verfahren zur Erkennung von Hass immer wichtiger. Da das einfache Löschen von hasserfüllten Beiträgen aber auch immer die Meinungsfreiheit einer Person bis zu einem gewissen Grad einschränkt sind andere Verfahren zum moderieren solcher Seiten gefragt.

Eine Strategie zur Bekämpfung von Hass ist die Entgegenwirkung von Hassrede mittels Gegenrede. Die Gegenrede generiert von Nutzenden ist ein wirksamen und ressourcen-schonendes Mittel, um Hass im Netz nachhaltig zu begegnen, da hier die Nutzenden selbst zu Moderierenden werden. Obwohl es zwar derzeit bereits Ansätze gibt Hass im Netz automatisch zu detektieren, gibt es noch wenig wissenschaftliche Arbeiten zur Erkennung von Gegenrede im Netz. Dies ist dahingehend wichtig, um Personen, die Gegenrede betreiben ausfindig zu machen, um sie in ihrer Sache zu bestärken. Somit könnte gefördert werden, dass diese Menschen noch mehr Gegenrede betreiben und gegebenenfalls auch andere Menschen dazu ermutigen selbst aktiv zu werden.
Ein Problem welches viele dieser Personen haben die Gegenrede betreiben ist, dass deren Kommentare oft selbst negativ behaftet sind (Gegenhass). Hier wäre ein weiterführender Gedanke, solchen Personen Wege und Möglichkeiten aufzuzeigen positive und sinnvollere Gegenrede zu erstellen.

Im Zuge der Recherche wurden einige englische, aber keine deutschen Datensätze zum Trainieren eines Maschine Learning-Modells gefunden das Gegenrede detektieren kann. Die Literatur ist sich zudem hinsichtlich der Strategie zum Erkennen von Gegenrede nicht einig. Überlegungen sind etwa, ob es sinnvoll ist den zu detektierenden Kommentaren auch einen Kontext, also den Ausgangsbeitrag mitzugeben um die Thematik dem Modell besser nahe zu bringen.
Ein weiteres Problem von großen und komplexen machine-learning Modellen wie etwa BERT-Transformern ist ihre Erklärbarkeit. Es ist nämlich oft nicht nachvollziehbar weshalb solche Modelle gewisse Entscheidungen treffen.

Der größte Fokus der Arbeit liegt daher darin, ein BERT-Modell für die Erkennung von Gegenrede zu trainieren und die Vorhersagen des Modells für den Menschen erklärbar zu machen. Derzeit gibt es nämlich keine wissenschaftlichen Arbeiten zur Erklärbarkeit eines Modells das es schafft Gegenrede im Internet zu erkennen. Auf Grund der fehlenden Datensätze wird der Fokus der Arbeit zunächst auf das trainieren eines englischsprachigen Modells gelegt, um mehr Ressourcen in die Recherche von geeigneten Strategien zum Trainieren und Erklären eines solchen Modells zu haben.
Im Zuge der Arbeit werden verschiedene Methoden zur Erklärbarkeit angewandt und miteinander verglichen. Neben bekannten Verfahren wie etwa Gradienten-basierten und Störungs-basierten Methoden, welche als Basis dienen sollen, werden auch zwei neuere Verfahren in die Arbeit mit einbezogen. Bei den neueren Methoden handelt es sich um eine Methode die auf den Aufmerksamkeits-Gewichten von dem verwendeten Transformer angelehnt ist, sowie eine Methode, die Prototypen für den zu erklärenden Input Satz erstellt. Zum Abschluss wird eine Evaluierung der Resultate erhoben um abzuklären welche der Verfahren am besten funktioniert, beziehungsweise mit welchen der Methoden Personen am besten zurechtkommen.

2 Status

2.1 **Meilenstein 1 –** Literaturrecherche (*fertig*)

Die Recherche wurde durch Anwendung verschiedener Suchmaschinen wie Google und Google Scholar durchgeführt. Auch Bibliotheken wie arXiv, IEEE, SpringerLink und paperswithcode wurden im Rahmen der Recherche verwendet.
13 Datensätze die herangezogen werden können umso ein Modell zu trainieren wurden zusammengetragen. Verschiedene Erklärbarkeitsmethoden mit Fokus auf Techniken die für Transformer Modelle und textuelle Daten verwendet werden können wurden gesammelt und deren Vor- und Nachteile festgestellt. Vorhandener Code für die gefundenen Erklärbarkeitsmethoden wurden auf GitHub gefunden sowie mögliche Python-Bibliotheken, welche diverse Verfahren implementiert haben wurden gesammelt.

Die Literaturrecherche ergab, dass es derzeit keine Ansätze zum Erkennen von deutscher Gegenrede und auch keine deutschen Datensätze für diese Aufgabe gibt. Zudem ist noch nicht ausreichend erforscht welche Trainingsstrategie die besten Resultate bringt. In vergangenen Werken wurden bereits einige Erklärbarkeitsmethoden auf BERT-Transformer angewandt, jedoch zeigen bekannte Methoden gewisse Einschränkungen. Gradienten-basierte Methoden werden in der Literatur häufig als unzuverlässig bezeichnet und nicht robust regenüber Veränderungen der Eingangsdaten. Störungsbasierte Methoden sind zudem häufig sehr langsam. Aufmerksamkeits-basierte Methoden sind zwar tendenziell schneller als Gradienten-basierte Methoden, jedoch

zeigen diese häufig nicht die Wichtigkeit eines Wortes, dass zu einer Entscheidung des Modells führt, sondern eher die Wichtigkeit des kontextuellen Embeddings des zu erklärenden Textes an. Gradienten-basierte Methoden werden zudem als besser geeignet bezeichnet.

Durch die breite des Themenfelds ist es oft nicht einfach verschiedene Arten der Erklärbarkeitsmethoden voneinander abzugrenzen. Einige davon haben in gewisser Weise auch Überschneidungen miteinander was das verstehen der einzelnen Ansätze schwierig macht. Eine große Aufgabe war es hier, den persönlichen Überblick über die Methoden nicht zu verlieren.

2.2 **Meilenstein 2 –** Auswahl von Erklärbarkeitsmethoden (*fertig)*

Durch die Literaturrecherche wurde ein Set an Erklärbarkeitsmethoden zusammengetragen von den vier verschiedenen Typen gewählt wurden um diese miteinander zu vergleichen. Es wurden Methoden basierend auf Gradienten, Perturbation, Attention und eine basierend auf Prototypen ausgewählt.
Gradienten dienen in Transformer Modellen zum updaten der Gewichtungen des Neuronalen Netzes. Bei der gewählten Methode (Integrated Gradients [1]) werden die Gradienten entlang des Pfades es Neuronalen Netzes Akkumuliert (von Basiswert bis bist zur Ausgabe). Damit kann ein Score berechnet werden (Attribution Score) der die Wichtigkeit eines Tokens (Einheit oder Sequenz von Zeichen in einem Text) spiegelt. Bei perturbation-basierten oder auch „störungsbasierten“ Methoden wird der Inputtext modifiziert und die Veränderung zum Output des Modells beobachtet umso die Wichtigkeit verschiedener Wörter in einem Text abzuschätzen. Hier wird die Methode LIME [2] gewählt wobei LIME eine lokale lineare Approximation zum Transformer Modell darstellt. Um diese Approximation zu erreichen wird eine Stichprobe aus den Daten gewählt und diese leicht verändert, um die Unterschiede zu den ursprünglichen Daten beobachten zu können. Für jeden Datenpunkt wird eine Gewichtung des Features berechnet. Mit Hilfe der Gewichtungen kann eine lokale Approximation des Transformers erstellt werden, mit der erklärt werden kann wie eine Vorhersage für einen bestimmten Bereich im Text getroffen wurde.
Die dritte Methode ist eine Attention-basierte Methode (GlobEnc [3]) bei der für die Vorhersage wichtige Wörter ebenfalls mit Hilfe eines Scores hervorgehoben werden. Der Attention-Mechanismus von Transformer Modellen wird verwendet, um Beziehungen innerhalb eines Textes zu modellieren und um die Aufmerksamkeit des Modelles auf bestimmte Teile oder Wörter im Text zu lenken. Für jedes Wort im Satz werden Vektoren erzeugt woraus Aufmerksamkeitsgewichte berechnet werden können. Um die Relevanz einzelner Wörter für die Vorhersage zu bestimmen werden die aus den Wörtern generierten Vektoren mit den Aufmerksamkeitsgewichte multipliziert.
Die letzte Methode ist eine Methode ist ProtoTEx [4] basierend auf so genannten Prototypen. Hier werden repräsentative Instanzen der Klassen in den Datenpunkten generiert. Diese Instanzen repräsentieren Merkmale und Charakteristiken der jeweiligen Klassen wieder. Diese Prototypen

dienen in der Vorhersage dann als Referenzwerte, die helfen zu erklären weshalb ein Modell eine Entscheidung getroffen hat.
Die vier Ansätze wurden gewählt, da sie sich in ihren Schwerpunkten unterscheiden und zueinander komplementär sind. Zusammen könnten Sie ein umfassenderes Verständnis für die Vorhersage von Transformer Modellen ermöglichen.

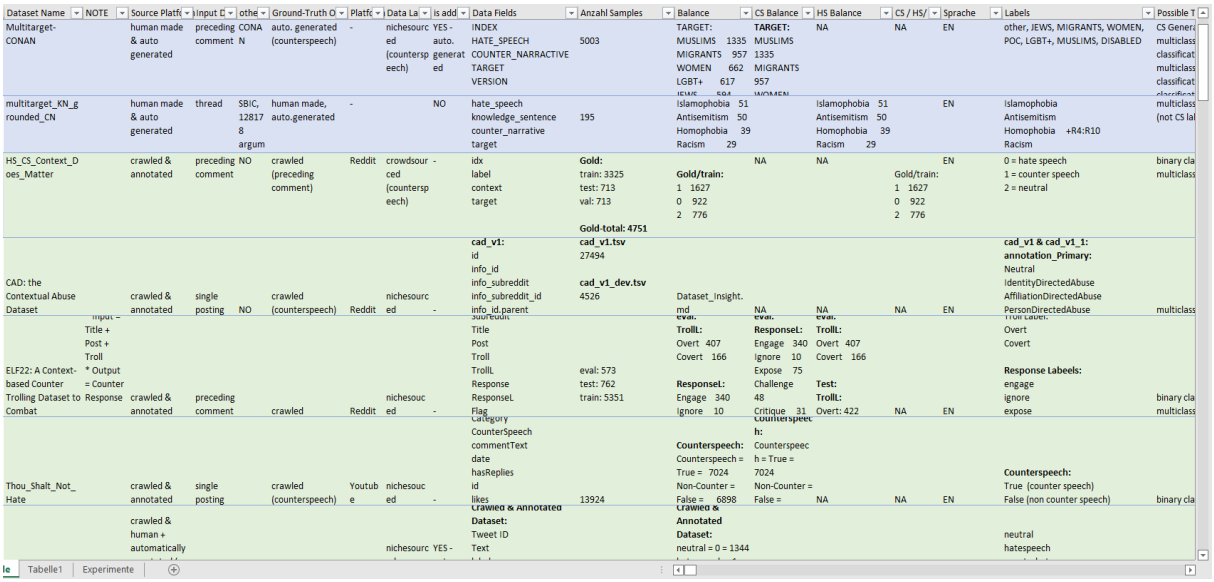
Diverse Python-Bibliotheken und Referenzcode wurden im Zuge dieses Meilensteins durchstöbert und potentielle Kandidaten ausgewählt.

2.3 Meilenstein 4 – Trainieren eines ersten Prototyps und Implementierung der Erklärbarkeitsverfahren *(noch in arbeit)*

Ein erstes BERT-Modell wurde trainiert wobei bei diesem noch kein Fokus auf eine gute Performanz gelegt wurde. Dieses Modell diente rein zum Testen der Erklärbarkeitsmethoden während deren Implementierung.
Drei der vier Erklärbarkeitsmethoden konnten bislang ohne weitere Probleme implementiert werden. Für den Ansatz basierend auf dem Attention-Mechanismus des Transformers war ursprünglich eine andere Methode als die nun verwendete geplant. Der Code der angedachten Methode ist leider nicht öffentlich zugänglich. Hier wurde deshalb per E-Mail bei dem Author nachgefragt mit der Antwort, dass der Code demnächst übermittelt wird. Auf diese Antwort wird bis Dato gewartet. Da seit Wochen nun keine Rückmeldung der Person zurück kam wurde eine alternative Methode (GlobEnc) gewählt. Die Implementierung der Prototyp-basierten Methode ist noch im Gange. Hier ist leider der Referenzcode nicht fehlerfrei. Eine weitere Hürde ist außerdem, das adaptieren des Codes für die Struktur des gewählten Transformer-Modelles. Den Code hierfür zum Laufen zu bringen stellt sich demnach als schwierig heraus.

2.4 Meilenstein 5 – Auswahl eines Datensets und trainieren eines BERT-Modells für Gegenrede *(noch in arbeit)*

Es wurden insgesamt 13 Datensets gefunden die sich dem Thema Gegenrede widmen. Einige der Datensets definieren Gegenrede als Gegenhass – diese negative Art von Gegenrede ist im Rahmen dieser Arbeit nicht gefragt, weshalb sich in der konkreten Auswahl für das zu trainierende Modell nur auf Datensätze mit positiver oder neutraler Gegenrede konzentriert wird. Um hier nicht nur einen geeigneten Datensatz zu wählen, sondern auch um zu Erkennen ob es sinnvoll ist den Ursprungsbeitrag mit einzubeziehen werden Experimente dahingehend durchgeführt. Eine ausführliche Tabelle mit den Details der gefundenen Datenset wurde erstellt. Ein kleiner Ausschnitt davon ist in Figure 1 Einblick in die entstandene Tabelle der Datenset-Recherche



Dataset Name	NOTE	Source Platfc	Input D	othe	Ground-Truth O	Platfc	Data La	is add	Data Fields	Anzahl Samples	Balance	CS Balance	HS Balance	CS / HS/	Sprache	Labels	Possible T
Multitarget-CONAN		human made & auto generated	preceding comment	CONA N	auto. generated (counterspeech)	-	nichesourc ed (counterspeech)	YES - auto. generated	INDEX HATE_SPEECH COUNTER_NARRATIVE TARGET VERSION	5003	TARGET: MUSLIMS 1335 MIGRANTS 957 WOMEN 662 LGBT+ 617	TARGET: MUSLIMS 1335 MIGRANTS 957 WOMEN	NA	NA	EN	other, JEWS, MIGRANTS, WOMEN, POC, LGBT+, MUSLIMS, DISABLED	CS Generi multiclass classificat multiclass classificat
multitarget_KN_grounded_CN		human made & auto generated	thread	SBIC, 128178 argum	human made, auto.generated	-		NO	hate_speech knowledge_sentence counter_narrative target	195	Islamophobia 51 Antisemitism 50 Homophobia 39 Racism 29		Islamophobia 51 Antisemitism 50 Homophobia 39 Racism 29		EN	Islamophobia Antisemitism Homophobia +R4:R10 Racism	multiclass (not CS la
HS_CS_Context_Does_Matter		crawled & annotated	preceding comment	NO	crawled (preceding comment)	Reddit	crowdsourced (counterspeech)	-	idx label context target	**Gold:** train: 3325 test: 713 val: 713 **Gold-total: 4751**	**Gold/train:** 1 1627 0 922 2 776	NA	NA	Gold/train: 1 1627 0 922 2 776	EN	0 = hate speech 1 = counter speech 2 = neutral	binary cla multiclass
CAD: the Contextual Abuse Dataset		crawled & annotated	single posting	NO	crawled (counterspeech)	Reddit	nichesourced	-	**cad_v1:** id info_id info_subreddit info_subreddit_id info_id.parent	**cad_v1.tsv** 27494 **cad_v1_dev.tsv** 4526	Dataset_Insight.md	NA	NA	NA	EN	**cad_v1 & cad_v1_1:** **annotation_Primary:** Neutral IdentityDirectedAbuse AffiliationDirectedAbuse PersonDirectedAbuse	multiclass
ELF22: A Context-based Counter Trolling Dataset to Combat	Input = Title + Post + Troll * Output = Counter Response	crawled & annotated	preceding comment		crawled	Reddit	nichesouced	-	Title Post Troll TrollL Response ResponseL Flag	eval: 573 test: 762 train: 5351	**TrollL:** Overt 407 Covert 166 **ResponseL:** Engage 340 Ignore 10 Expose 75 Challenge 48 Critique 31	**ResponseL:** Engage 340 Ignore 10 Expose 75 Challenge 48 Critique 31	**TrollL:** Overt 407 Covert 166 **Test:** **TrollL:** Overt: 422	NA	EN	Overt Covert **Response Labeels:** engage ignore expose	binary cla multiclass
Thou_Shalt_Not_Hate		crawled & annotated	single posting		crawled (counterspeech)	Youtube	nichesouced	-	Category CounterSpeech commentText date hasReplies id likes	**Counterspeech:** Counterspeech = True = 7024 Non-Counter = False = 6898	**Counterspeech:** Counterspeech = True = 7024 Non-Counter = False =	NA	NA	NA	EN	**Counterspeech:** True (counter speech) False (non counter speech)	binary cla
		crawled & human + automatically					nichesourc	YES -	Tweet ID Text	**Dataset:** neutral = 0 = 1344						neutral hatespeech	

Figure 1 Einblick in die entstandene Tabelle der Datenset-Recherche

Die am besten geeignete Methode wird gewählt und ein BERT-Modell mit einem passenden Datensatz trainiert. Ebenfalls werden die Trainingsparameter des Modells so angepasst, dass eine möglichst gute Performanz erzielt wird.

2.5 Meilenstein 6 – Anwendung der Erklärbarkeitsmethoden auf das fertige Modell (kommender Meilenstein)

Sobald ein gutes funktionierendes Modell trainiert wurde werden die Erklärungsmethoden auf eines der Vorhersagen des Modelles angewendet.

2.6 Meilenstein 7 – Evaluierung der Erklärbarkeitsmodelle und Diskussion der Ergebnisse (kommender Meilenstein)

Zumindest drei der gewählten Methoden liefern am Ende so genannte Erklärbarkeits-Scores, welche die Wichtigkeit der jeweiligen Wörter im Text anzeigen die zur Entscheidung des Modells beitragen. Diese können wiederrum verwendet werden um eine statistische sowie eine Evaluierung durch Menschen durchzuführen. Im Rahmen der Evaluierung soll erkannt werden, welche Unterschiede die Methoden im direkten Vergleich mit sich bringen und ob sie Menschen tatsächlich helfen die Entscheidungen des Modells besser zu verstehen.

3 Zusammenfassung Planaktualisierung

Alle Anpassungen des Planungsdokuments kurz zusammengefasst

Im Planungsdokument wurden die einzelnen Meilensteine noch etwas konkreter dargestellt. Der grundlegende Plan hat sich jedoch nicht verändert.

[1] J. D. Janizek, P. Sturmfels, und S.-I. Lee, „Explaining Explanations: Axiomatic Feature Interactions for Deep Networks“, *J Mach Learn Res*, Bd. 22, Nr. 1, Juli 2022.
[2] M. T. Ribeiro, S. Singh, und C. Guestrin, „‚Why Should I Trust You?‘: Explaining the Predictions of Any Classifier“. arXiv, 9. August 2016. doi: 10.48550/arXiv.1602.04938.
[3] A. Modarressi, M. Fayyaz, Y. Yaghoobzadeh, und M. T. Pilehvar, „GlobEnc: Quantifying Global Token Attribution by Incorporating the Whole Encoder Layer in Transformers“, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Juli 2022, S. 258–271. doi: 10.18653/v1/2022.naacl-main.19.
[4] A. Das, C. Gupta, V. Kovatchev, M. Lease, und J. J. Li, „ProtoTEx: Explaining Model Decisions with Prototype Tensors“. arXiv, 22. Mai 2022. doi: 10.48550/arXiv.2204.05426.