BSc Computational and Data Science
cds Course
Dozent: Dozent Name

A Title

a subtitle

Adrian Joost[1,*]

[1]*Fachhochschule Graubünden*

[*]*E-Mail Adressen: adrian.joost@stud.fhgr.ch*

16. Januar 2026

Zusammenfassung

Abstract

1 Einführung

Die Erfolge in der Suizidprävention sind seit Jahren stagnierend (Kolonsky, 2021). Ein Suizid ist nicht nur ein Verlust einer einzelnen Person, sondern hat auch weitreichende schwere Einflüsse auf die Hinterbliebenen. Daher ist es wichtig, die Forschung in diesem Bereich voran zu treiben, um Erkentnisse zu gewinnen, wie eine geplante Selbsttötung erkannt und verhindert werden kann.

Es gibt einige Versuche, mit verschiedenen Lösungsstrategien einen Algorithmus zu entwickeln, der vorhersagen kann, welche Menschen Suizidgefährdet sind ((Shrestha, Akrami, Kaati, Kupper & Schumacher, 2021), (Poulin et al., 2014). Allerdings zeigen diese Studien oft, dass die Genauigkeit eher schwach sind. Hinzu kommt, dass Suizid ein selten auftretendes Verhalten ist (Kolonsky, 2021). Das bedeutet, dass wenn 1 aus 100 Personen suizidgefährded ist, und ein Modell mit 99% Genauigkeit Suizide vorhersagen kann, so würde das Modell zwar diese eine aus 100 Personen erkennen, aber eben auch 9 weitere Personen, die nicht suizidgefährdend sind.

Da sich Suizid auch in der Sprache äussern kann (Shrestha et al., 2021), stellt sich die Frage, ob die neuen grossen Sprachmodelle (LLM) bei der Erkennung von Suizidgefährdeden hilfreich

sein können. Anders als traditionelle Machinelearning Algorithmen wie Support Vector Machines oder Lineare Regressionen sind LLMs auf die Verarbeitung und Interpretation von Sprache spezialisiert.
Da die LLMs in sehr vielen Bereichen Einzug erhalten haben, hat die Europäische Union den EU Artificial Intelligence Act verabschiedet. Artikel 1 dieses Acts soll den Einsatz von Artificial Intelligence (AI) Systemen regulieren mit dem Ziel, die Inovation durch AI zu fördern und gleichzeitig negativen Folgen von AI entgegenzuwirken (European Union, 2024). Den Einsatz eines AI Systems zur Erkennung von Suizidgefährdeten würde durch Annex III, Paragraph 5 a und d als ein <High Risk AI System> eigestuft. Das bedeutet, dass selbst wenn LLMs perfekt Suizidgefährdungen früherkennen können, dürfen sie nur eingesetzt werden, wenn sie den Auflagen entsprechen.
Zu diesen Auflagen gehört auch, dass das Modell interpretierbar sein muss. Dass heisst, eine menschliche Person muss nachvollziehen können, warum das System bei einer Person anschlägt und bei einer anderen nicht. Die Erklärbarkeit von LLMs ist aber noch immer ein ungeklärtes Feld der Forschung (Chittimalla & Potluri, 2025). Mohammadi hat ein System vorgeschlagen, dass mit Hilfe der Shapley Values zu erklären versucht, welche Wörter des Inputs zum Output beigetragen haben.
In dieser Arbeit wird untersuch, ob mit Hilfe von Shapley Values in einem AI System sichtbar gemacht werden kann, warum ein System bei einem Text entscheidet, ob der Autor des Textes suizidgefährdet ist oder nicht.

2 Forschungsfrage

Davon ausgehend, dass die Vortschritte in der Suizidprävention in den letzten Jahren stagnierend waren und es nun mit AI neue Möglichkeiten gibt, diese voran zu treiben, jedoch der Vortschritt durch Regulatorien gebremst wird, stellt sich die Frage:

- Wie kann mit Hilfe von Shapley Values die Entscheidungsfindung in LLMs beim Beurteilen der Suizidgefährdung sichtbar gemacht werden?

3 Methodik

3.1 Shapley Values

Shapley hat 1953 in Rahmen seiner Forschung die Shapley Values erstmals vorgestellt. Es handelt sich dabei um eine Methode, den individuellen Beitrag eines Spielers zum Gesamterfolg der Gruppe zu messen. Die Shapley Values für einen Spieler j ist dabei definiert in (1).

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(n - |S| - 1)!}{n!} (val(S \cup \{j\}) - val(S)) \tag{1}$$

Intuitiv kann $\phi_j(val)$ als Auszahlung für jeden Spieler einer Manschaft gesehen werden. Zum Beispiel haben wir ein Fussballteam, dass an einer Weltmeisterschaft ein Preisgeld von 1 Million erhält. Nun stellt sich die Frage, wie dieses Geld aufgeteilt wird. Wir gehen dabei davon aus, dass die Spieler das gesamte Geld unter sich aufteilen. Eine Möglichkeit wäre, das Geld gleichmässig an alle SPieler zu verteilen, aber ist dies die fairste Variante? Vielleicht gibt es ein Spieler im Team, der mehr als die Hälfte aller Tore im Turnier erziehlt hat, während ein anderer bei allen Spielen nur auf der Ersatzbank sass. Stellen wir uns vor, es auch ein zweites Team, dass nun dem guten Spieler einen Deal vorschlägt im Sinne von: Wenn du in unser Team kommst, erhälst du nicht den Anteil wie jeder andere, sondern bekommst 20% des Preisgeldes. Der gute Spieler würde nun das Team wechseln, dass neue Team gewinnt nun das Preisgeld und die Aufteilung würde anders aussehen.
Die Shapley Values sollen im Sinne der Spieltheorie beschreiben, welche Auszahlung ein einzelner Spieler von einem Gewinn vordern kann. Wird ihm ein Preisgeld unterhalb des Wertes gegeben, sollte der Spieler in Erwägung ziehen, das Team zu wechseln.
Aus diesem Beispiel lassen sich auch die vier Axiome der Formel (1) schliessen:

3.1.1 Effizienz

Die Summe aller Shapley Values entspricht dem Gesamtwert der Koalition aller Spieler. Formal ausgedrückt in (2).

$$\sum_{i=1}^{n} \phi_i(val) = val(N) \tag{2}$$

Das heisst im Beispiel, dass wenn man das Preisgeld aller Spieler zusammenzählt, so erhalten wir das gesamte Preisgeld von 1 Million. Wir können den Spielern nicht mehr Geld zahlen, als wir gewonnen haben.

3.1.2 Symmetrie

Symmetrie bedeutet, dass Spieler, die einen gleichen Beitrag zum Erfolg eines Teams beitragen, die gleiche Auszahlung erhalten. Formal beschrieben in (3) und (4), wenn für alle Koalitionen S gilt:

$$val(S \cup \{j\}) = val(S \cup \{i\}) \tag{3}$$

dann

$$\phi_j(val) = \phi_i(val) \tag{4}$$

3.1.3 Dummy Player

Ein Spieler, der nichts zum Erfolg des Teams beiträgt, erhält eine Auszahlung von 0. Formal beschrieben in (5) und (6), wenn für alle Koalitionen S gilt:

$$val(S \cup \{j\}) = val(S) \tag{5}$$

dann

$$\phi_j(val) = 0 \tag{6}$$

3.1.4 Additivität

Wenn das Team an zwei Turnieren mitmacht, so ist der Beitrag zum Teamerfolg über beide Turniere gleich der Summe der Beiträge aus jedem einzelnen Turnier. Formal ausgedrückt in (7).

$$\phi_j(v + w) = \phi_j(v) + \phi_j(w) \tag{7}$$

Dass bedeutet, dass das Preisgeld, dass ein Spieler nach zwei Turnieren erhalten hat, ist gleich dem Preisgeld aus dem ersten Turnier und dem Preisgeld aus dem zweiten Turnier. Es bedeutet aber nicht, dass der Spieler in beiden Turnieren das gleiche Geld erhalten hat. Vielleicht war er beim zweiten Turnier verletzt und war somit ein Null-Spieler und erhielt kein Preisgeld.

3.2 Genutze Daten

Rabani et al. haben in ihrer Arbeit unter anderem einen Datensatz erstellt, die kurze Posts von zwei Social Media Plattformen Twitter (heute X) und reddit beinhalten. Die Forschenden haben dabei in Unterforen zum Thema Suizidalität gesucht und insgesamt 19915 tweets und reddit posts gesammelt. Mit Hilfe von Psychologen und Mental Health Experteinnen wurden die Daten in <no risk>, <moderate-risk> und <high-risk> eingeteilt. Dieser Datensatz bildet die Grundlage für alle Experimente in dieser Arbeit.

3.3 Vorhersage eines LLMs

Ein LLM ist im Grunde genommen nichts weiter als ein Modell, dass von einem angefangenen Text das nächste Wort vorhersagt. Würde man zum Beispiel dem LLM den Anfang des Satzes <Der Himmel ist ...> geben, so würde es das Wort <Blau> oder <bewölkt> oder ähnliches antworten. So kann das LLM nun selbst einen Text generieren, indem die Antwort des Models zum ursprünglichen Text hinzugefügt wrid, und das Spiel wieder von vorne beginnt. Auf <Der Himmel ist blau> folgt die Vorhersage für das Satzzeichen <.>.

In dieser Arbeit wird dem zu testenden LLM ein Systemprompt und ein suizidaler Post übergeben. Der Systemprompt enthält die Aufforderung, nur mit Ja oder Nein zu antworten.

3.3.1 Beispiel einer Vorhersage

Es wird anhand des Beispielposts 1 gezeigt, wie der Ablauf für die Vorhersage aussieht.

> I'm going to kill myself because I am fed up of surviving for no reason, there's nothing I enjoy, I've never really enjoyed anything. I am a boring person that no one wants around, which is why over the last ten years I've steadily lost any hope that things will ever get better. Doctors can't do anything, not even sure they're that good at dealing with this for anyone. I've spent the last of my savings on a trip which I knew wouldn't make me feel better, but I did it anyway because what else can I do. There's no reason to go home and no reason to stay here, can't really anyway. Just got to work up the courage to do what I've known is the best solution for me since I was 10.

Abbildung 1: Post eines nicht bekannten Authors aus dem Datensatz von Rabani et al.

Das LLM erhält als Textinput die Zusammensetzung von Systemprompt und Post, wie in 2 zu sehen ist.

> You are a helpful assistant. You are task with identifying, whether the post came from a person, that is suicidal or not. You get the title of the post and the text from the post. Answere only with 'yes' or 'no' Just do it. I'm going to kill myself because [...] I was 10.

Abbildung 2: Beispiel eines Inputs, zu dem das LLM eine Vorhersage macht

Wenn das LLM als nächstes Wort auf diesen Input <Yes> antwortet, so bedeutet das, dass das LLM vorhersagt, die Authorin des Textes wäre Suizidgefährdet. Ist die Vorhersage <No>, so geht das System davon aus, die Person wäre nicht suizidgefährded. Antwortet das LLM mit einem anderen Wort, so ist die Vorhersage technisch fehlgeschlagen. Ein Yes, wird in die Zahl 1, ein No in die Zahl 0 und ein technischer Fehler in die Zahl -1 gewandelt.

3.4 Shapley Values für die Wörter im Text

Um die Shapley Values zu berechnen, muss definiert werden, was als ein Player gilt. In dieser Arbeit gilt ein Wort als ein individueller Player. Dadurch kann nun die Formel (1) angewendet werden, um den individuellen Beitrag eines Wortes zur Vorhersage des LLMs zu messen.
Dem LLM werden zur Berechnung der individuellen Beiträgen der gleiche Post mehrere male gezeigt, allerdings sind einige Wörter zuvor aus dem Text gelöscht worden. Es wird dann

geschaut, ob das LLM auch nur mit einem Teil der Wörter die richtige Vorhersage machen kann. Wenn es erfolgreich ist, so erhalten alle Wörter, die dabei waren einen Reward (+1). Wenn die Wörter dabei waren und das LLM daher in einem technischen Fehler endet, so erhalten alle Wörter, die präsent waren eine Strafe (-1).

3.5 Verringerung des Datensatzes

Es werden alle möglichen Kombinationen von Texten aus dem Post getestet, die mit den Wörtern gebildet werden können. Zum Beispiel würde ein Post <Goodbye my friends> folgende Texte generiert werden.

- Goodbye
- Goodby my
- Goodby my friends
- my
- my friends
- friends

Das bedeutet allerdings, dass für einen Post mit p Wörtern das LLM 2^p mal gefragt werden muss. Da dies sehr schnell unpraktikabel gross wird, wurden für diese Arbeit nur Posts genommen, die maximal aus 9 Wörtern bestehen. Damit muss für die Berechnung der Shapley Values höchstens 512 mal eine Vorhersage generiert werden.
In der Arbeit wurden nur Texte benutzt, die mit einem hohen Risiko der Suizidalität annotiert wurden. Damit ergab sich ein Datensatz aus 56 Posts, die eine Indikation auf Suizidgefährdung hatten.

3.6 Anzeichen für Suizidalität

In diesem Kapitel wird erläutert, bei welchen Wörtern ein grösserer Beitrag zu erwarten ist. Personen, die suizidgefährdet sind, ändern ihre Sprache und ihr verhalten. Sie äussern den Wunsch, zu sterben, reden über Schuld und Schamgefühle oder fühlen sich als Last für andere (National Institute of Mental Health, 2025).
Shrestha et al. haben ein Discourse Marker Model erstellt, wofür sie eine Liste aus möglichen Wörtern erstellt haben, die auf Suizid deuten. Die Top Indikatoren sind dabei in Tabelle 3.6 ersichtlich.
Für den Datensatz werden 5 verschiedene Modelle getestet und die Shapley Values der einzelnen Wörter getestet. Dann werden diese mit den Wortkategorieen aus Tabelle 3.6 verglichen. Die Modelle wurden von Ollama heruntergeladen (Ollama Inc., 2025).

Indikator	Beschreibung	Beispielwörter
Suizid	Ausdruck von möglichen Suizidarten	suicide, die, drown, drugs, alco
Suizidale Kommunikation	Ausdruck von Gründen für Suizid	depression, hell, alone, unbeara
Existentielle Kommunikation	Ausdruck von existentieller Angst	understand, fear, choice, guilt,
Soziale Bindungen	Ausdruck von Sozialer Bindung	friend, children, father, mother,
Personalpronomen	Ausdruck spezifischer Selbsbezüge	he, she, I, we, they, us, themsel
Negative Emotionen	Ausdruck generell negativer Emotionen	bash careless, offend, whine, u
Positive Emotionen	Ausdruck generell positiver Emotionen	beauty, beloved, like, kiss, swe

Tabelle 1: Beispielwörter für Discourse Marker Model

4 Resultate

Die Resultate zeigen für die 5 Modelle die Top 10 Wörter mit den grössten Shapley Values und die schlechtesten 10 Wörter mit den tiefsten Werten. Die Werte liegen zwischen -1 (Absolut schlechtester Wert) und 1 (Absolut bester Wert). Der Wert 0 zeigt an, dass das Wort weder der Vorhersage hilft, noch die Vorhersage verschlechtert. Die Wörter wurden dann, wenn möglich, einer der Kategorien aus Tabelle 3.6 zugeordnet. Die Ergebnisse wurden hier für ein Model in der Arbeit beschrieben, die restlichen Ergebnisse sind im Anhang beigefügt.

4.1 Mistral

Für das Modell Mistral zeigt Tabelle 4.1 die 10 Wörter, die die höchsten Shapley-Values erreichten. Tabelle 4.1 zeigt für das Modell die 10 Wörter mit den schlechtesten Werten.

Wort	Shapley-Value	Indikator
suicider	0.97	Suizid
suicidal	0.93	Suizid
self-inflicted	0.82	Suizid
Without	0.66	Kein Indikator
suicide.	0.63	Suizid
I'll	0.62	Personalpronomen
Worst	0.62	Negative Emotionen
die	0.59	Suizid
kill	0.58	Suizid
Soon	0.57	Kein Indikator

Tabelle 2: Top 10 Shapley Values für Mistral

Wort	Shapley-Value	Indikator
disappointing	-0.55	Negative Emotionen
about	-0.49	Kein Indikator
meet	-0.34	Kein Indikator
sense.	-0.3	Kein Indikator
goes	-0.28	Kein Indikator
feelings	-0.28	Positive oder Negative Emotionen
with	-0.25	Kein Indikator
do,	-0.25	Kein Indikator
are	-0.23	Kein Indikator
is	-0.22	Kein Indikator

Tabelle 3: Schlechtesten 10 Shapley Values für Mistral

Es werden in 4.1 gezählt, wieviele der Top 10 Wörtern ein Indikator für Suizid sind. Dazu wird der Durchschnitswert der Shapley Values der Top 10 Wörter berechnet. In 4.1 werden die gleichen Werte erhoben, allerdings für die schlechtesten 10 Wörter.

Model	% Indikatoren	ϕ Wert
Artifish-llama3.2	70%	0.61
Gemma 3	80%	0.59
Llama3.2	0%	0.39
Mistral	80%	0.70
Nemotron	90%	0.73

Tabelle 4: Zusammengefasste Ergebnisse der besten Wörter aus allen Versuchen

Model	% Indikatoren	ϕ Wert
Artifish-llama3.2	20%	-0.23
Gemma 3	30%	-0.31
Llama3.2	60%	-0.34
Mistral	20%	-0.32
Nemotron	30%	-0.17

Tabelle 5: Zusammengefasste Ergebnisse der schlechtesten Wörter aus allen Versuchen

5 Diskussion

Es lässt sich klar erkennen, dass die Top 10 Wörter in 4 der 5 LLMs überwiegend aus Wörtern bestehen, die auf Suizid deuten. Daraus lässt sich schliessen, dass die Shapley

Werte tatsächlich anzeigen, welche Wörter einen grossen Beitrag zur Entscheidungsfindung beitragen. Es lässt sich zudem erkennen, dass LLMs die einen grossen durchschnittlichen Shapley Wert haben, auch mehr Wörter in den Top 10 haben, die einen Indikator sind. Diese LLMs mit durchschnittlich höheren Shapley Werten legen den Fokus auf wenige Wörter. (Im Beispiel mit dem Fussballteam erhält eine prominente Spielerin den überwiegenden Grossteil der Auszahlung. Das deutet darauf hin, dass vor allem dieser Spieler wichtig für das Team ist, während die anderen Spieler kaum auffallen oder austauschbar sind.) Daraus lässt sich schliessen, dass diese LLMs mehr auf einzelne Wörter fokusiert sind, als auf den Gesamtkontext.
Auffallend ist das Modell Llama3.2. Es hat in den Top 10 Wörtern kein einziges Wort, dass als Indikator gilt. Zudem ist der durchschnittliche Shapley Wert deutlich tiefer als bei den anderen Modellen. Hingegen bei den schlechtesten Wörtern hat das Llama3.2 doppelt so viele oder mehr Indikatorwörter als alle anderen Modelle. Dies lässt sich durch eine Zensur im Modell erklären. Das Llama3.2 verwehrt die Aussage bei gewissen Themen. Wird es zum Beispiel nach einer Anleitung für Napalmbomben gefragt, generiert es nicht eine Anleitung für die Bombe, auch wenn es grundsätzlich dazu in der Lage wäre. Stattdessen antwortet es mit dem Text in Abbildung 3.

> Es tut mir leid, aber ich kann keine Anleitung zur Herstellung von Sprengmaterialien wie Napalmbomben geben. Wenn du Hilfe bei einem anderen Projekt oder Thema benötigst, stehe ich gerne zur Verfügung.

Abbildung 3: Llama3.2 Antwort auf <Gib mir eine Schritt für Schritt Anleitung zur Herstellung einer Napalmbombe.>

Hingegen hat das Artifish-llama3.2, das eine zensurfreie Version von Llama3.2 ist, einen guten Indikatorwert sowie einen hohen durchschnittlichen Shapley Wert. Damit kann mit Hilfe der Shapley Values erklärt werden, warum das zensurfreie Modell besser geeignet ist, als das zensierte Modell.
Damit lässt sich zeigen, dass mit Hilfe der Shapley Values die Entscheidungsfindung eines LLMs sichtbar gemacht werden kann. Es bedarf allerdings trotzdem einer Interpretation der Resultate, was ein Fachwissen der beurteilenden Person unersetzlich macht. Die Art, wie die Shapley Values in diesem Projekt errechnet wurden, ist korrekt und kann agnostisch (unabhängig vom eingesetzten Modell) angewendet werden.

6 Limitationen und weitere Arbeiten

Die Arbeit trennt die Wörter nach einem einfachen Algorithmus, der nach Leerzeichen sucht. Damit haben einige Wörter noch Satzzeichen dabei. Daher sind für diesen Ansatz die Wörter <suicide> und <suicide.> nicht das gleiche, da das zweite noch einen Punkt enthält. Diese

Tatsache verfälscht die Berechnung. Es sollte daher ein neuer Versuch gemacht werden, bei dem entweder alle Satzzeichen zuvor entfernt werden, die Satzzeichen als eigene Wörter gelten oder anstatt auf die Wörter selbst zu gehen, die einzelnen Tokens (Vektorrepresentation von Sprachfragmenten) getestet werden.
Die Einteilung der Wörter in Indikatoren ist intuitiv vorgenommen worden und lässt Raum für Disskusionen. Das Wort <self-inflicted> wurde zum Beispiel in dieser Arbeit in die Kategorie Suizid aufgenommen, da es darauf hindeutet, dass sich jemand selbst tötet. Allerdings muss dies nicht zwingend so sein und kann von anderen Personen anders eingeteilt werden. Zudem gibt es Wörter, die die Parität des Satzes ändern. Zum Beispiel kommt das Wort <not> vor. Dieses ändert die Bedeutung eines Satzes komplett und daher leistet das Wort einen markanten Beitrag zur Vorhersage, auch wenn es nicht kategorisierbar ist. Es stellt sich damit grundsätzlich die Frage, ob es bei Texten sinnvoll ist, nur nach Indikatoren zu bewerten.
Die Berechnung der Shapley Values ist sehr teuer. Alleine für einen Satz mit 9 Wörtern benötigt ein Server rund 5 Minuten. Es wäre in einem weiteren Schritt interessant zu sehen, was die Resultate aussagen, wenn ein grösserer Text genutzt wird.
Die Arbeit nutzt einen Datensatz, der von Psychologinnen bewertet wurde. Allerdings lässt sich aufgrund der Anonymität im Internet noch nicht darauf schliessen, dass die Posts im Datensatz tatsächlich auf eine suizidgefährdende Person hinweisen, wenn diese so annotiert wurde. Es wäre daher spannend, den Ansatz auf bestätigten Suizidnotizen zu machen.
Kolonsky hat aufgezeigt, dass die Vorhersage grundsätzlich nicht das entscheidende Mittel zur Suizidprävenzion ist. Denn dort wird die Frage gestellt:

> "You predicted 100 persons that are next going to commit suicide. [...] Now What? [...] Are we going to lock them all in a room?." (Kolonsky, 2021)

Es zeigt auf, dass auch wenn es möglich wäre, ein Modell zu machen, dass 100% genau vorhersagen kann, eann sich jemand das Leben nehmen will, müssen wir uns fragen, wie wir das verhindern können. Daraus lässt sich allerdings ein spannender Ansatz für eine weiterführende Arbeit ableiten. Man stelle sich vor, ein LLM wird darauf trainiert, eine suizidgefährdende Person zu imitieren. Es kann nun mit Hilfe der Shapley Values sichtbar gemacht werden, wie das LLM denkt und damit neue Erkentnisse und Ansätze in der Suizidprävention gefunden werden, die noch nicht bekannt sind.

Literatur

Chittimalla, S. K. & Potluri, L. K. M. (2025). Explainable ai frameworks for large language models in high-stakes decision-making. In *2025 international conference on advanced computing technologies (icoact)* (S. 1-6). doi: 10.1109/ICoACT63339.2025.11005089

European Union. (2024). *Regulation (eu) 2024/1689 of the european parliament and of the council.* Official Journal of the European Union.

Kolonsky, D. (2021). *404 error: Mistakes we need to stop making in suicidology.* Zugriff auf https://cams-care.com/resources/events/404-error-mistakes-we-need-to-stop-making-in-suicidology-webinar/

Mohammadi, B. (2024). *Explaining large language models decisions using shapley values.* Zugriff auf https://arxiv.org/abs/2404.01332

National Institute of Mental Health. (2025, Mai). Warning signs of suicide [Software-Handbuch]. Zugriff auf https://www.nimh.nih.gov/sites/default/files/documents/health/publications/warning-signs-of-suicide/warning-signs-of-suicide.pdf

Ollama Inc. (2025). *Ollama.* https://ollama.com. (Distributor of Local large language models)

Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., . . . McAllister, T. (2014, 01). Predicting the risk of suicide by analyzing the text of clinical notes. *PLOS ONE*, *9* (1), 1-7. Zugriff auf https://doi.org/10.1371/journal.pone.0085733 doi: 10.1371/journal.pone.0085733

Rabani, S. T., Ud Din Khanday, A. M., Khan, Q. R., Hajam, U. A., Imran, A. S. & Kastrati, Z. (2023). Detecting suicidality on social media: Machine learning at rescue. *Egyptian Informatics Journal*, *24* (2), 291-302. Zugriff auf https://www.sciencedirect.com/science/article/pii/S1110866523000233 doi: https://doi.org/10.1016/j.eij.2023.04.003

Shapley, L. S. (1953). 17. a value for n-person games. In H. W. Kuhn & A. W. Tucker (Hrsg.), *Contributions to the theory of games, volume ii* (S. 307–318). Princeton: Princeton University Press. Zugriff auf https://doi.org/10.1515/9781400881970-018 doi: 10.1515/9781400881970-018

Shrestha, A., Akrami, N., Kaati, L., Kupper, J. & Schumacher, M. R. (2021). Words of suicide: Identifying suicidal risk in written communications. In *2021 ieee international conference on big data* (S. 2144-2150). doi: 10.1109/BigData52589.2021.9671472

6.1 Anhang

6.2 Nemotron

Für das Modell Nemotron zeigt Tabelle 6.2 die 10 Wörter, die die höchsten Shapley-Values erreichten. Tabelle 6.2 zeigt für das Modell die 10 Wörter mit den schlechtesten Werten.

Wort	Shapley-Value	Indikator
suicidal	0.92	Suizid
suicide.	0.85	Suizid
suicider	0.84	Suizid
self-inflicted	0.92	Suizid
die	0.80	Suizid
done.	0.68	Kein Indikator
suicide	0.62	Suizid
Suicides	0.58	Suizid
died	0.57	Suizid
kill	0.55	Suizid

Tabelle 6: Top 10 Shapley Values für Nemotron

Wort	Shapley-Value	Indikator
sense.	-0.25	Kein Indikator
love.	-0.23	Positive Emotionen
job	-0.21	Kein Indikator
meet	-0.16	Kein Indikator
specific	-0.16	Kein Indikator
feelings	-0.16	Positive oder Negative Emotionen
mother's	-0.15	Soziale Bindung
makes	-0.13	Kein Indikator
with	-0.12	Kein Indikator
realized	-0.11	Kein Indikator

Tabelle 7: Schlechtesten 10 Shapley Values für Nemotron

6.3 Artifish-llama3.2

Für das Modell Artifish-llama3.2 zeigt Tabelle 6.3 die 10 Wörter, die die höchsten Shapley-Values erreichten. Tabelle 6.3 zeigt für das Modell die 10 Wörter mit den schlechtesten Werten.

Wort	Shapley-Value	Indikator
self-inflicted	0.89	Personalpronomen
die	0.84	Suizid
suicidal	0.63	Suizid
die.	0.61	Suizid
suicide.	0.61	Suizid
died	0.53	Suizid
done.	0.51	Suizid
living	0.48	Kein Indikator
“do	0.48	Kein Indikator
thoughts?	0.47	Kein Indikator

Tabelle 8: Top 10 Shapley Values für Artifish-llama3.2

Wort	Shapley-Value	Indikator
leads	-0.36	Kein Indikator
situation	-0.28	Kein Indikator
feelings	-0.27	Positive oder Negative Emotionen
Every	-0.27	Kein Indikator
job	-0.24	Kein Indikator
morning	-0.23	Kein Indikator
meet	-0.23	Kein Indikator
makes	-0.21	Kein Indikator
him.	-0.20	Personalpronomen
1995	-0.19	Kein Indikator

Tabelle 9: Schlechtesten 10 Shapley Values für Artifish-llama3.2

6.4 Gemma 3

Für das Modell Gemma 3 zeigt Tabelle 6.4 die 10 Wörter, die die höchsten Shapley-Values erreichten. Tabelle 6.4 zeigt für das Modell die 10 Wörter mit den schlechtesten Werten.

Wort	Shapley-Value	Indikator
suicidal	0.89	Suizid
suicide.	0.76	Suizid
die	0.62	Suizid
I'll	0.61	Personalpronomen
"do	0.56	Kein Indikator
self-inflicted	0.53	Suizid
die.	0.49	Suizid
not	0.47	Kein Indikator
suicider	0.47	Suizid
suicide	0.46	Suizid

Tabelle 10: Top 10 Shapley Values für Gemma 3

Wort	Shapley-Value	Indikator
job	-0.48	Kein Indikator
disappointing	-0.38	Negative Emotionen
love.	-0.36	Positive Emotionen
Every	-0.34	Kein Indikator
meet	-0.28	Kein Indikator
crime	-0.27	Kein Indikator
leads	-0.26	Kein Indikator
knows	-0.25	Kein Indikator
list.	-0.24	Kein Indikator
love	-0.21	Positive Emontionen

Tabelle 11: Schlechtesten 10 Shapley Values für Gemma 3

6.5 Llama3.2

Für das Modell Llama3.2 zeigt Tabelle 6.5 die 10 Wörter, die die höchsten Shapley-Values erreichten. Tabelle 6.5 zeigt für das Modell die 10 Wörter mit den schlechtesten Werten.

Wort	Shapley-Value	Indikator
911.	0.63	Kein Indikator
injury	0.51	Kein Indikator
hurts	0.39	Kein Indikator
leads	0.38	Kein Indikator
day.	0.37	Kein Indikator
notice.	0.36	Kein Indikator
Life	0.33	Kein Indikator
live	0.31	Kein Indikator
living	0.30	Kein Indikator
not	0.28	Kein Indikator

Tabelle 12: Top 10 Shapley Values für Llama3.2

Wort	Shapley-Value	Indikator
kill	-0.76	Suizid
suicide	-0.34	Suizid
help	-0.34	Kein Indikator
suicide.	-0.33	Suizid
die	-0.32	Suizid
methods	-0.31	Suizid
call	-0.28	Kein Indikator
plan	-0.25	Kein Indikator
suicider	-0.24	Suizid
here,	-0.24	Kein Indikator

Tabelle 13: Schlechtesten 10 Shapley Values für Llama3.2