Richtlinie zum „Software-Sprint“

KIZervus - KI zur Erkennung vulgärer Sprache

Schlussbericht

Zuwendungsempfänger:

Nicolas Kuhaupt

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS22S07 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Vulgäre Sprache im Sinne von unangemessen wird hier als Obermenge von sexistischer, rassistischer oder beleidigender Sprache verstanden. Diese tritt nicht nur auf Social Media Plattformen auf, sondern auch in Foren (z.B. Nachrichtenportale) oder in Profilen (z.B. Datingportale). Um vulgärer Sprache entgegenzuwirken, werden bisher Kommentarfunktionen für sensible Nachrichtenrubriken geschlossen oder Arbeitnehmer:innen überprüfen Texte manuell.

Durch eine auf KI basierende Erkennung kann ein Arbeitsschritt teilautomatisiert werden. Dadurch können Diskussionen geführt werden ohne vulgäre Sprache befürchten zu müssen.

Dieses Projekt wurde in drei Schritten umgesetzt:

1. Datensammlung

Für das Training von KI-Modellen werden große Mengen an Daten benötigt. Diese wurden zusammengetragen und in ein einheitliches Format überführt werden.

2. Training des KI-Modells

Das Training beziehungsweise Finetuning eines KI-Modells funktioniert nicht stringent nach einem Plan, sondern ist mit Experimenten verbunden. Für diesen Meilenstein wurden Experimente durchgeführt und ein Modell gefunden, das eine gute Performance bei der Klassifizierung vulgärer Sprache liefert.

3. Implementierung als Service

Im finalen Schritt wurde das Modell in einen Docker Container eingebunden, der eine REST Schnittstelle bietet. Außerdem wurde eine technische Anleitung geschrieben, damit das KI-Modell mit den eigenen Daten angepasst und in die eigene Plattform integriert werden kann.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Themenfeld des Softwaresprints war unter anderem Hate Speech. Während Hate Speech per Definition gegen Menschen gerichtet ist, umfasst die vulgäre Sprache auch Texte, die unangemessen formuliert sind und sich nicht zwangsweise gegen Menschen wendet. Damit ist vulgäre Sprache das Ausschlussziel jedes Plattformbetreibers. Adressiert werden hier jegliche online Angebote, die Veröffentlichung von User geschriebenen Text ermöglichen (Kommentarfunktionen, Chats, Profilseiten etc.) Betreiber von diesen Plattformen müssen Inhalte filtern und bei entsprechendem Inhalt löschen. Dies ist häufig noch manuelle Arbeit und kann mit einer KI unterstützt bzw. zum aktuellen Zeitpunkt teilautomatisiert werden. Wenige (große) Plattformen haben eigene Lösungen, bieten aber weder das KI-Modell, noch die zugrundeliegenden Daten öffentlich an. Daher ist hier auch der Open Source Ansatz wichtig. Auch in der Klassifizierung von Texten sind Biase und Diskriminierung präsent. Hier wird das KI-Modell inklusive Datenbasis offen zur Verfügung gestellt. Dadurch wird die viel kritisierte Intransparenz deutlich gesenkt. Gleichzeitig kann das bestehende Modell um eigene Daten ergänzt und auf eigene Bedürfnisse angepasst werden.

Ausführliche Darstellung der Ergebnisse

In diesem Projekt wurde der größte bekannte Open Source Datensatz für vulgäre Sprache und Hate Speech zusammengetragen. Jeder andere Datensatz, der bereits Open Source war, wurde entsprechend in das Projekt integriert. Der Datensatz stammt aus diversen Quellen (Social Media, Nachrichtenkommentare und Chatnachrichten) und ist thematisch breit aufgestellt. Hierdurch sind insgesamt etwa 97000 klassifizierte Texte zusammengetragen worden.

Das trainierte Modell zeigt eine Performance von >80% Richtigkeit, d.h. 80% aller Texte werden korrekt als vulgär bzw. nicht vulgär klassifiziert. Der zugrundeliegende Datensatz besteht (durch Data Augmentation) zur Hälfte aus vulgären und nicht-vulgären Texten. Wird der Datensatz eingegrenzt und homogener gestaltet (beispielsweise aus einer der oben genannten Quellen), wird eine Performance von ~90% erreicht. Daher kann realistisch bei Finetuning mit eigenen Daten mit einer Performance von 90% gerechnet werden.

Damit das Finetuning bzw. die Adaption auf den eigenen Use Case und die Erweiterung auf die eigenen Daten stattfinden kann, ist der Trainingsprozess gut dokumentiert und einfach gestaltet. Die Daten müssen in das einheitliche Format überführt werden und das Training kann mit wenigen Schritten beginnen. Das Training findet dabei mit Parametern statt, die sich als robust erwiesen haben.

Im nächsten Schritt kann das Modell einfach deployed werden. Dafür existiert die Vorlage und eine Integration als REST Schnittstelle ist in wenigen Schritten möglich.

Ursprünglich wurde eine bessere Performance erwartet. Die Ergebnisse lassen keine vollständige Automatisierung der Spracherkennung zu. Nichtsdestotrotz kann das KIZervus Modell ein Baustein sein, der Inhalte vorfiltert. Anschließend müssen nur noch Texte mit hoher Unsicherheit der Klassifizierung (das Modell kann zusätzlich zur vorhergesagten Klasse einen Wahrscheinlichkeitsscore ausgeben) manuell überprüft werden.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Zielgruppe sind auf der einen Seite Betreiber von online-Plattformen, die Texte von Usern öffentlich machen, z.B. Nachrichtenportale, Social Media Seiten, oder Webseiten mit Chatfunktion. Das KIZervus-Repository kann als Grundlage für das eigene Modell genutzt werden. Hier sind zwei Dinge wertvoll. Erstens der Datensatz, der als Grundlage oder Ergänzung dienen kann. Es zeigt sich, dass KI Modelle besser funktionieren, je mehr Daten sie zum Training zur Verfügung haben. Zweitens ist ein Trainingsprozess implementiert, der schnell adaptiert werden kann und robust gute Ergebnisse liefert. Weiterhin kann das Modell anschließend mit wenigen Schritten produktiv gestellt werden.

Auf der anderen Seite kann das Repository für weitere Experimente genutzt werden. Wie bereits aufgezeigt, bietet das Modell bisher eine Performance von etwa 80% Richtigkeit. Aktuell gibt es schnelle Fortschritte im Bereich der Sprachverarbeitung. Mit dem Repository können schnell neue Experimente durchgeführt werden. Aber auch die Erweiterung durch neue Daten ist problemlos und schnell möglich.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Wie bereits erwähnt, wurde ursprünglich eine bessere Performance des Modells erwartet. Gleichzeitig deckt sich die Performance mit anderen Arbeiten, die eine Hate Speech Klassifizierung durchgeführt haben. Bei der Evaluierung des KIZervus Modells wurde insbesondere auf diejenigen Texte geschaut, die eigentlich vulgär sind, von dem Modell aber nicht als solche klassifiziert wurden. Hier wurden oft Metaphern, Vergleiche und Ironie angetroffen. Nach Einschätzung des aktuellen State of the Art ist es nicht möglich, diese zielsicher automatisiert zu erkennen. Es bleibt abzuwarten, ob dort bessere Sprachmodelle oder gar neue Techniken kommen werden.

Weiterhin wurde anfangs geplant, die Datensammlung durch eigens gescrapte Daten zu bereichern. Damit sind Daten gemeint, die auf Plattformen anfallen und automatisiert heruntergeladen und gespeichert werden. Diese müssten dann noch eigens als vulgär oder nicht vulgär gelabelt werden. Hier haben sich zwei Hindernisse bzw. Gründe dagegen gezeigt. Zum einen ist das Labeln der Daten nicht einfach. Dazu bestehen auch Best Practices, die unter anderem mindestens zwei Personen zum Labeln vorsehen. Damit soll die Qualität der Labels verbessert werden. Zum anderen war es deutlich schneller, bereits gelabelte Daten zu finden und zusammengetragen. Diese Daten bildeten auch eine ausreichende Grundlage für das Training des Modells.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Eine Demonstration der Klassifikation ist unter folgendem Link erreichbar:
https://huggingface.co/KIZervus/KIZervus

Das Github Repository ist hier erreichbar: https://github.com/NKDataConv/KIZervus

Bei Interesse kann eine Mail an nk[at]data-convolution.de geschrieben werden.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Ein Großteil des Projektes entfiel auf das Training des KI Modells. Wie bereits mehrfach erwähnt, wurde hier eine bessere Performance erwartet, sodass darauf viel Zeit verwendet wurde. Gleichzeitig kann dabei beliebig viel und beliebig lange experimentiert werden und stets eine bessere

Performance erreicht werden. Dadurch konnte der Arbeitsplan und die Zeit gut gemanagt werden. Während der Projektlauftzeit wurde der Meilenstein 3 (Deployment des Modells) vorgezogen, um sicherzustellen, dass das Projekt rund ist. Dann wurde zum Training zurückgekehrt und die Performance weiter verbessert.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Sprachverarbeitung ist im Moment ein wichtiges und verbreitetes Thema. Im speziellen handelt es sich um Large Language Modells (LLM). Das sind KI Modelle, die Sprache verarbeiten und groß im Sinne von vielen Neuronen sind. In der Forschung gibt es aktuell ständig neue Durchbrüche. Auch Startups mit LLMs sprießen aus dem Boden. Interessant und im Sinne dieses Projektes ironisch ist beispielsweise OpenAIs Copilot. Dies ist eine Autovervollständigung beim Schreiben von Code. Copilot geht über die klassische Autovervollständigung weit hinaus und kann ganze Code Passagen schreiben. Copilot basiert ebenfalls auf den Technologien, die hier im Projekt genutzt wurden (also LLMs). Ein auf LLM basierendes Modell hat also dabei unterstützt, ein anderes LLM zu programmieren.