mVISE

Toxicity Detection

Entwicklung eines Modells zur Erkennung von toxischen Kommentaren in englischsprachigen Online-Konversationen.

Zusammenfassung

Die Data Science Experten der mVISE AG haben ein effizientes Modell für die vorurteilsfreie, semantische Textanalyse entwickelt. Damit kann „*Hate Speech*“ besser erkannt werden, was sowohl für die Nutzer als auch die Betreiber sozialer Plattformen von Vorteil ist.

Kontext

„*Hate Speech*“ im Internet ist ein zunehmendes Problem für soziale Plattformen. Die Nicht-Erkennung toxischer Kommentare birgt die Gefahr, dass Benutzer verbal angegriffen werden und die Plattformbetreiber strafrechtlich verfolgt werden. Die Blockierung falsch erkannter Hassrede dagegen schränkt die freie Meinungsäußerung ein und führt zu einer negativen Benutzererfahrung. Facebook steht bei diesem Thema an vorderster Front, aber die Auswirkungen sind selbst für Inhaber kleiner Chat-Foren oder jedes Portals, welches textbasierte Chats erlaubt, relevant.

Herausforderung

Die Betreiber sozialer Plattformen sind dazu verpflichtet, rechtswidrige Inhalte innerhalb eines bestimmten Zeitraums nach Eingang einer Beschwerde zu löschen. Diese Haftung führt zu erheblichen Kosten für die Überprüfung von Textinhalten. Selbst wenn leistungsfähige KI-Lösungen eingesetzt werden, bleibt die Rate der falsch klassifizierten Kommentare hoch, sodass eine manuelle Überprüfung der Inhalte erforderlich ist. Die Komplexität der Sprache und ihre Interpretation - also, dass Wörter mehrere Bedeutungen haben, die auf subjektiven Kontexten und Einstellungen basieren - erfordern eine semantische Analyse, die äußerst schwer zu programmieren ist.

> ***“Wenn ich eine perfekte KI hätte, um Inhalte – Texte, Fotos und Videos – zu moderieren, würde das unser Leben beträchtlich erleichtern.”***
>
> Jerome Pesenti // FaceBook VP of Artificial Intelligence

Die Lösung

mVISE-Experten nahmen an einer *Open Innovation Challenge* auf der *Kaggle*-Website teil, um „*Toxizität in einer Vielzahl von Online-Konversationen zu erkennen* “. Mit einem Datenset aus etwa 100.000 gekennzeichneten Kommentaren, die in Trainings- und Testdatensätze unterteilt waren, arbeitete das Team iterativ an der Entwicklung von Algorithmen zur Erkennung toxischer Kommentare. In Anlehnung an Scrum-Arbeitsmethoden lieferte das Entwicklungsteam regelmäßig Software-Uploads und erhielt von der Kaggle-Website Bewertungen, die auf der Genauigkeit, Sensitivität und Spezifität der Lösung basierten. Der Ansatz bestand in der Verwendung eines *Natural Language Processing* (NLP)-Toolsets namens BERT (*Bidirectional Encoder Representations from Transformers*), das auf der Basis von öffentlich zugänglichem Online-Text vortrainiert ist und eine bidirektionale Analyse des Textkontextes ermöglicht. Im Vergleich zu anderen Analysen verbraucht dieser Ansatz mehr Ressourcen, führt aber zu einer größeren analytischen Genauigkeit.

Vorgehensweise

Die Lösung wurde mit den folgenden Tools und Technologien entwickelt:

- **Python:**
  Als Programmiersprache für den selbstlernenden Algorithmus
- **BERT:**
  Ein vortrainiertes Open-Source-Tool für bidirektionale Textanalysen
- **Google Colab:**
  Eine Cloud-basierte Entwicklungsumgebung für das Training des Machine Learning-Modells

mVISE-Experten waren innerhalb kürzester Zeit in der Lage, ca. 90% der Kommentare aus den Testdatensätzen richtig als toxisch oder nicht toxisch zu klassifizieren.

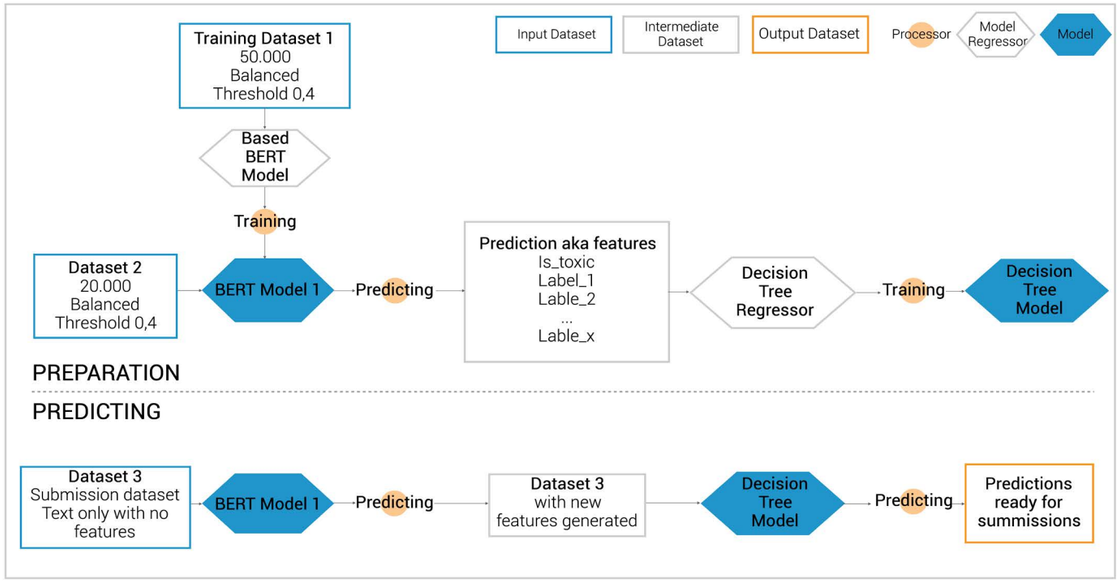

Abb. 1 Vorgehensmodell der Toxicity Detection anhand der Trainingsdatensätze

Empfehlungen

Besitzer von sozialen Plattformen sollten automatisierte Methoden zur Erkennung von toxischen Kommentaren und Hassreden bevorzugen. Es zeigt sich, dass der händische Aufwand zur Erkennung von „*Hate Speech*“ durch immer fortschrittlichere Algorithmen bedeutend eingedämmt werden kann. Dabei ist zu beachten, dass universell anwendbare Algorithmen schwer zu finden sind und sozialen Plattformen angeraten wird, plattform- und inhaltsspezifische Lösungen zu verwenden. Vorgeschulte Open-Source-Tools, wie BERT, sollten zur Beschleunigung der Entwicklung und zur Verringerung des Bedarfs an umfangreichen Trainingsdatensätzen verwendet werden. Für das erfolgreiche Training der Modelle empfiehlt es sich dennoch, eine möglichst große Menge an spezifischen Daten bereitzustellen.

Haben Sie ebenfalls Interesse an einer individuellen, datengetriebenen Lösung für Ihr Unternehmen? Sie suchen einen Partner zur Beratung oder auch Umsetzung Ihres Data Science Projekts?

Dann sprechen Sie uns an. Die Experten-Teams der mVISE beraten Sie gern.
service@mvise.de | **www.mvise.de**

mVISE

Wir unterstützen mittelständische und große Unternehmen aller Branchen dabei, von der digitalen Revolution zu profitieren. Die besondere Kombination aus firmeneigenen Software-Lösungen mit ausgewählten Experten-Teams in den relevanten und aktuellen IT-Themengebieten schafft nachhaltige Wettbewerbsvorteile für unsere Kunden.

Unsere Experten bestimmen, gestalten, kreieren und steuern IT-Infrastrukturen und Software-Lösungen für Datenintegrations- und Enterprise-Data-Management-Projekte, mit dem Ziel, die aktuellen Geschäftsmodelle unserer Kunden zukunftssicher zu machen und gleichzeitig neue Geschäftsmodelle zu identifizieren.

Sprechen Sie uns an – gerne stellen wir Ihnen unser Angebot
in einem persönlichen Gespräch näher vor.
service@mvise.de | www.mvise.de

mVISE AG
Wahler Straße 2
40472 Düsseldorf
Fon: +49 211 78 17 80 – 0