Projektskizze Semesterprojekt CAS Big Data

Titel: Aufbau einer Streaming und Analyse Umgebung für detect-hate.com

1. Umfeld / Ausgangslage

Im CAS Practical Machine Learning wurde ein erster Durchstich für eine automatisierte Erkennung von Hasskommentaren (hate speech) in Twitter Tweets erstellt. Als Datengrundlage für die Modelle dienten drei unterschiedliche Datensätze aus verschiedenen Quellen mit total ca. 80'000 Tweets. Diese Datensätze enthielten nebst den eigentlichen Tweets weitere Spalten, so z.B. die folgenden Labels:

- None
- Abusive
- Sexist
- Racist

Basierend auf den verwendeten Datensätzen und deren Labels wurden unterschiedliche Klassifikationsalgorithmen, wie auch ein Neuronales Netz trainiert, um auch neuen, den Klassifikationen unbekannte Hasskommentare automatisiert zu erkennen. Die trainierten Algorithmen wurden in eine Web-Applikation (mit Python Flask realisiert) integriert, welche aktuell auf einem privaten NAS läuft und nach wie vor unter www.detect-hate.com aufrufbar ist.

2. Problemstellung

Tests mit neuen, in der Applikation eingegebenen Texten von Twitter (oder auch frei erfunden) haben gezeigt, dass die trainierten Algorithmen nur bedingt korrekte Aussagen treffen können. Dieses Problem liegt insbesondere in den für das Training verwendeten Datensätzen:

1. Zu wenig Trainingsdaten für eine Textklassifikation von Freitext, insbesondere für das Neuronale Netz
2. Die Tweets der Trainingsdaten wurden jeweils nur über einen sehr kurzen Zeitraum gesammelt, das führt zu einem grossen Bias Problem (zu dem Erfassungszeitraum behandelte Themen entsprechen nur z.T. den Themen über einen längeren Zeitraum resp. zur aktuellen Zeit)
3. Nicht alle Datensätze hatten Labels für alle Kategorien von Hasskommentaren (weniger als 10’000 sexistisch und/oder rassistisch gelabelte Daten).

Um diese Probleme zu lösen sollen Daten einerseits direkt via Twitter API gesammelt und in eine r geeigneten Datenbank gespeichert werden und andererseits soll der Zeitraum der Sammlung länger andauern (über das Abgabedatum dieser Arbeit hinaus).

Durch die direkte Sammlung von Twitter sind die so gespeicherten Tweets nicht aufbereitet und enthalten keine Labels, allerdings können Hashtags, zum Beispiel #everydaysexism als Labels verwendet werden.

Ziel dieser Arbeit ist das Speichern grosser Mengen von Daten direkt ab Twitter, welche einerseits direkt mittels Stream Analytics in Dashboards angezeigt werden können, andererseits dass die Daten längerfristig für weiteres Training der Algorithmen aus der vorangegangenen Semesterarbeit verwendet werden können.

3. Lösungsansatz

Es soll eine Streaming und Analyse Umgebung aufgebaut werden. Die Tweets sollen per Stream direkt von der Twitter API bezogen werden. Anschliessend erfolgen erste pre-processing Schritte. Die bereinigten Tweets werden zum Einen für ein Dashboard zur “Stimmungsanzeige” genutzt, zum Anderen für weitere Bereinigungsschritte und anschliessendes Trainieren der Machine Learning Modelle in einer geeigneten Form persistiert werden.

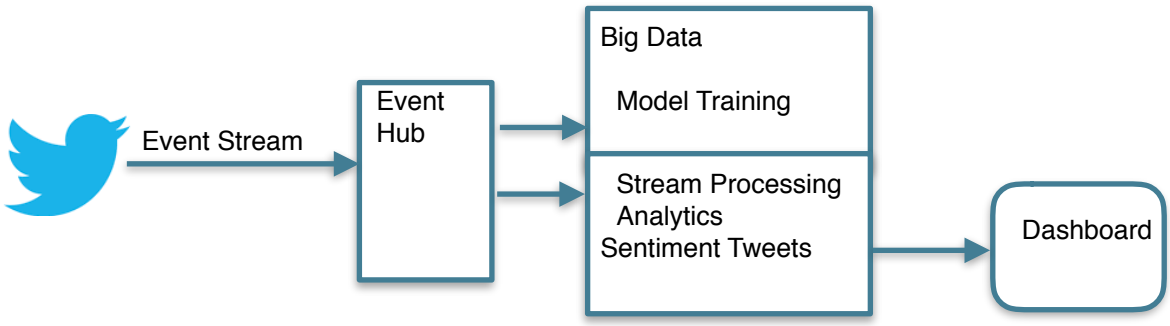

Abbildung 1: Aufbau Streaming und Analyse Umgebung

Der Aufbau der Streaming und Analyse Umgebung soll (sofern praktikabel) on-prem aufgebaut werden. Alternativ kommt ein Cloud Ansatz in Frage.

Spezifikation des BigBoards Cluster “nHex i3”:

- 6 nodes
- Intel NUC
- X86_64
- Intel Core i3ß
- 12 cores, 24 threads
- 96 GB RAM
- 6 TB storage

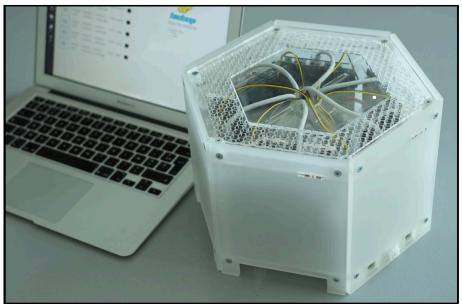

Abbildung 2: nHex i3

4. Personen

Studierende: SBB, Mai, Verena, 079759 71 36, verenamai@gmail.com