Räumliche Verortung von textbasierten Social-Media-Einträgen am Beispiel von Polizei-Tweets

Svenja Ruthmann, Alexander Rolwes, Klaus Böhm

Im Rahmen dieses Beitrags wird die räumliche Verortung deutscher Tweets auf Basis von verfügbaren Standardwerkzeugen untersucht.

Hintergrund

Die Forschungsinitiative untersucht Möglichkeiten zur Verortung von textbasierten Social-Media-Einträgen mit verfügbaren Bibliotheken und Diensten. Der Fokus liegt hierbei auf deutschen Kurznachrichten (Tweets) des Mikrobloggingdienstes Twitter. Die Bedeutung der Verarbeitung speziell deutscher Sprache wird seit der steigenden Nutzung von Twitter in der öffentlichen Verwaltung zunehmend relevanter. Allein die Polizei verwaltet im Jahr 2017 schon mehr als 200 Accounts auf Twitter und Facebook (Anzlinger, 2019). In diesem Zusammenhang werden die Kurznachrichten häufig von Büropersonal verfasst, sodass die in Tweets optional enthaltene Standortangabe keine Information im Zusammenhang mit dem textlichen Inhalt liefert.

Ansatz

Der methodische Ansatz für das Vorhaben ist wie folgt: Zunächst werden die Herausforderungen durch eine Literaturrecherche sowie durch einen vorverarbeitenden Schritt zur Identifizierung besonderer charakteristischer Merkmale in einem Tweet adressiert. Anschließend folgt die Definition eines algorithmischen Ablaufes aus den gewonnenen Erkenntnissen. Die Auswahl geeigneter und verfügbarer Werkzeuge liefert die Basis für die prototypische Umsetzung. Eine Evaluation der Ergebnisse bewertet die Untersuchung.

Als spezielle Herausforderung bei der Verortung von deutschen Tweets zeigt sich insbesondere die maximale Zeichenlänge jener von 280 Zeichen. Eine im November 2018 veröffentlichte Studie benennt die durchschnittliche Länge eines Tweets mit lediglich 33 Zeichen (AFP, 2018). Um mit dieser Einschränkung eine Vielzahl an Informationen zu teilen, ist es üblich in den verfassten Kurznachrichten die Grammatik des Textes zu vernachlässigen. Zusätzlich werden auch Emoticons und Abkürzungen verwendet, um die Kurznachrichten mit der ge-

Abb. 1: Tweet der Polizei Rheinpfalz

wünschten Information anzureichern. Eine weitere Schwierigkeit besteht darin, dass häufig auch Wörter der englischen Sprache in eine deutsche Struktur eingebettet werden. Zudem sind die Ortsangaben oft unkonkret formuliert. Problematisch wird es v. a. dann, wenn ein in Deutschland mehrfach existierender Stadtname erwähnt wird. Des Weiteren werden Städtenamen teilweise nicht vollends ausgeschrieben, sondern als Abkürzung angegeben – bspw. in Anlehnung an die deutschen KFZ-Kennzeichen. In Abb. 1 ist exemplarisch ein typischer Tweet einer Polizeistation dargestellt.

Der entwickelte algorithmische Ablauf lässt sich wie folgt gliedern: In einem vorverarbeitenden Schritt werden Umlaute ersetzt sowie Sonderzeichen, Emoticons und Links entfernt. Für die Ermittlung der Ortsnamen unterteilt der Natural Language Prozessor spaCy (Explosion AI, 2019) die Tweets in Entitäten. Um zu überprüfen, ob eine Abkürzung eines KFZ-Kennzeichens enthalten ist, werden alle Eigennamen mit einer Datenbank abgeglichen und ggf. im Tweet ersetzt. Anschließend wird der Geocoder HERE (HERE Global B.V., 2019) eingesetzt, um die Koordinaten zu generieren. Alle ohne Ortsbezug bestehenden Eigennamen werden ergebnislos zurückgegeben und nicht weiterverfolgt.

Im Rahmen der Auswahl geeigneter Werkzeuge wurden Alternativen gegenübergestellt. Bei einem Vergleich verschiedener Natural Language Prozessoren wurden kommerzielle Anbieter ausgeschlossen, da diese keinen Einblick in den Ablauf gewähren und zudem nicht konfiguriert werden können. Eine zentrale Anforderung für die Sprachanalyse ist der sichere Umgang mit der deutschen Sprache sowie der Umgang mit den speziellen Eigenschaften von Tweets. Bei einem zu verwendenden Geocoder galt die Untersuchung sowohl kommerzieller Produkte als auch Open Source Software. Als erfolgreiche Open Source Alternative zur Ermittlung von Koordinaten aus Ortsnamen wurde OSMNames (Klokan Technologies GmbH, 2016) erkannt. Der Nachteil daran ist, dass die Namen in englischer Sprache abgespeichert sind und die deutsche Bezeichnung nur in den alternativen Namen vermerkt ist. Mit dem kommerziellen Anbieter HERE lässt sich eine große Anzahl an Abfragen kostenlos durchführen und wurde deshalb OSMNames gegenübergestellt. Aufgrund der Ortsnamen in deutscher Sprache lassen sich mit HERE bessere Ergebnisse erzielen, sodass dieser Dienst eingesetzt wird.
Die Anwendung des beschriebenen Algorithmus erzeugt eine JSON-Struktur welche beispielhaft für den o. g. Tweet dargestellt ist. Der beschriebene Prozess wurde in einem Prototyp implementiert. Der Nutzer wählt in einer Webanwendung eine Polizeibehörde aus, deren Tweets verortet werden sollen. Das Programm bezieht die Tweets mit Hilfe der Twitter API (Twitter Inc., 2019) und analysiert diese nach o. g. Verfahren. Die Ergebnisdarstellung folgt auf einer Karte.


```
{"text": "Stadtblitzer heute in Edigheim, Maudach und
      Pfingstweide. Ludwigshafen‘“,
"date": "2019-08-14 04:06:00",
"author": "Polizei Rheinpfalz",
"location": "Edigheim Ludwigshafen",
"locLabel": "Edigheim, Ludwigshafen am Rhein, Rheinland-
      Pfalz, Deutschland",
"latitude": 49.53106,
"longitude": 8.38829}
```

Evaluation

Zur Evaluation wurde ein Test mit 100 Tweets von verschiedenen Polizeibehörden durchgeführt, welche eine unterschiedliche Anzahl an Ortsangaben beinhalten. Dabei fand eine Analyse der Tweets sowohl mit der entstandenen Software als auch manuell statt. Die Untersuchung betrachtete zunächst den Natural Language Prozessor sowie den Geocoder getrennt voneinander, um anschließend das Ergebnis nach dem Ablauf des gesamten Programms zu prüfen. Da spaCy den Text in Entitäten unterteilt (Explosion AI, 2019), folgte eine Zuordnung dieser in die Kategorien „Ortsname“, „kein Ortsname“ und „Ortsname nicht erkannt“. Innerhalb der 100 Tweets ermittelte spaCy 91 Ortsnamen und 118 sonstige Eigennamen – 19 Ortsangaben wurden nicht erkannt (Siehe Abb. 2a). Bei einer Übermittlung lediglich vollständiger Ortsnamen an den Geocoder ordnete dieser 82 Namen richtig, 13 nicht erkennbar und fünf falsch zu (siehe Abb. 2b). Das Ergebnis des gesamten Programmablaufs betrug 68 korrekt ermittelte, fünf falsche und 29 nicht erkannte Orte (siehe Abb. 2c).

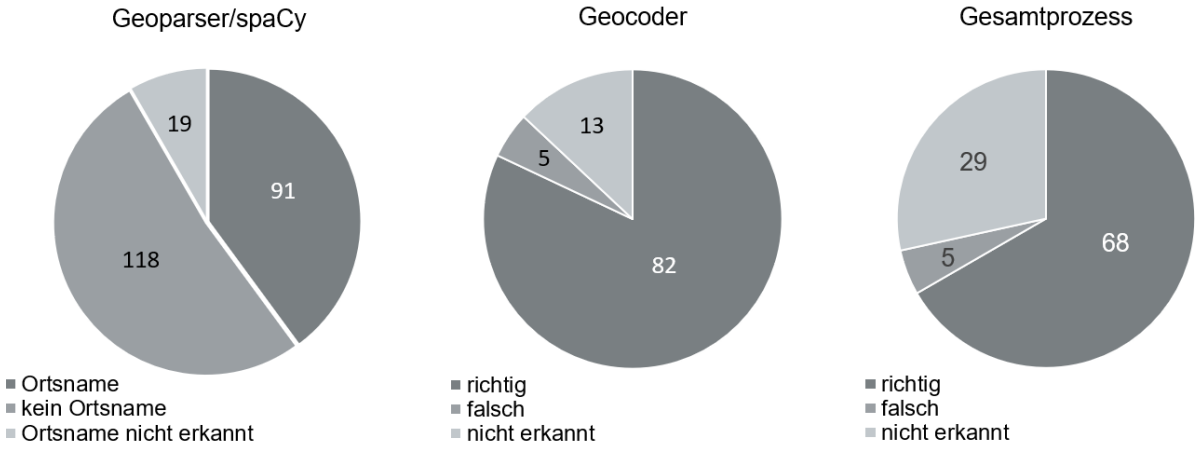

Abb. 2: Testergebnis

In den Tweets werden außerdem Himmelsrichtungen genutzt, um einen Bereich innerhalb der Stadt einzugrenzen. Diese werden vom Algorithmus ebenfalls erkannt, obwohl nicht alle Städte einen danach benannten Stadtteil besitzen. In dem Fall kann die Ortsangabe nur auf die Genauigkeit der Stadt bestimmt werden. Zudem waren Autobahnen, Kreise und Bundesländer nicht hinterlegt. Bei Hinzunahme dieser wird die Anzahl nicht erkannter Ortsangaben signifikant sinken. Die größte Verbesserung des gesamten Prozesses ließe sich erreichen, wenn spaCy mit einem Trainingsdatensatz deutscher Polizei-Tweets trainiert wird. Dadurch können die Entitäten direkt nach dem Attribut „Location“ abgefragt werden (Explosion AI, 2019), sodass keine anderen Ortsnamen zu fehlerhaften Ergebnissen führen. Darüber hinaus wäre eine vollständigere Ermittlung der Ortsangabe möglich.

Das Ergebnis der Untersuchung legt dar, dass verfügbare Bibliotheken und Dienste eine solide Basis für die Verortung von Tweets liefern können. Sie führen jedoch nicht zu einem akzeptablen Niveau, weshalb eine Prozesskette mit weiteren Zwischenschritten entwickelt werden muss. Exemplarisch dient bei dieser Untersuchung der eingesetzte Abgleich von Städtenamen als KFZ-Abkürzung sowie die spezielle Behandlung von mehrdeutigen Ortsangaben (z. B. Frankfurt) durch Nutzung der räumlichen Distanz zur twitternden Stelle.

Kontakt zum Autor:

Svenja Ruthmann	Alexander Rolwes	Klaus Böhm
Hochschule Mainz	Hochschule Mainz	Hochschule Mainz
Lucy-Hillebrand-Straße 2	Lucy-Hillebrand-Straße 2	Lucy-Hillebrand-Straße 2
55128 Mainz	55128 Mainz	55128 Mainz
svenja.ruthmann@gmx.de	alexander.rolwes@hs-mainz.de	klaus.boehm@hs-mainz.de

Literatur

AFP. (31. Oktober 2018). Ein Jahr nach der Verdopplung der Twitter-Textlänge sind die Tweets... kürzer.

Von Stern: https://www.stern.de/news/ein-jahr-nach-der-verdopplung-der-twitter-textlaenge-sind-die-tweets----kuerzer-8426332.html

Anzlinger, J. (13. November 2019). Hier spricht die @Polizei. Von Süddeutsche Zeitung:

https://www.sueddeutsche.de/panorama/staatsgewalt-auf-social-media-hier-spricht-die-polizei-1.3645986 abgerufen

Explosion AI. (27. Juli 2019). spaCy. Von https://spacy.io/ abgerufen

HERE Global B.V. (27. Juli 2019). Developer. Von HERE: https://developer.here.com/ abgerufen

Klokan Technologies GmbH. (2016). OSM Names. Von https://osmnames.org/docs/

Twitter Inc. (27. Juli 2019). Developer. Von Twitter: https://developer.twitter.com/ abgerufen