Lokalisierung von Tweets

Bachelor-Thesis von Johannes Nachtwey
August 2012

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Informatik
Knowledge Engineering

SAP RESEARCH

Betreuer:	Prof. Dr. Johannes Fürnkranz
Verantwortliche Mitarbeiter:	Dr. Heiko Paulheim (TU Darmstadt)
	Dipl.-Wirt. Inform. Axel Schulz (SAP Research)

Vorgelegte Bachelor-Thesis von Johannes Nachtwey

1. Gutachten:
2. Gutachten:

Tag der Einreichung:

Erklärung zur Bachelor-Thesis

Hiermit versichere ich, die vorliegende Bachelor-Thesis ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 22. August 2012

(Johannes Nachtwey)

Zusammenfassung

Soziale Medien, wie Twitter, generieren eine große Menge an Daten. Täglich werden Millionen an Tweets versendet, welche diverse verwertbare Informationen enthalten können. Um eine Nutzung der Informationen im Katastrophenfall zu ermöglichen, ist es essentiell, einen geografischen Bezug zu den Nachrichten herzustellen. Aktuell sind nur knapp ein Prozent aller Tweets mit einem Koordinatenpaar versehen. Deshalb besteht die Herausforderung darin, ein geeignetes Verfahren zu finden, Tweets geografisch zu lokalisieren. Mit der in dieser Arbeit vorgestellten Methode, Tweets auf Basis des Standortfelds zu verorten, ist es möglich, diese präziser zu lokalisieren als mit den bisher bekannten Verfahren.
Der vorgestellte Ansatz besteht darin, den geografischen Eintrag im Standortfeld als Approximation der Position des Tweets zu verwenden. Um die Präzision der Prognose der Tweetposition nachhaltig zu erhöhen, wurden weitere Verbesserungen in Form von Filtern präsentiert. Dabei ist es möglich, Länderangaben zu filtern, die Suche auf Städte einzugrenzen und ausschließlich Koordinaten aus dem Standortfeld zu betrachten sowie die einzelnen Filter zu kombinieren. Weitere Informationen aus dem Nutzerprofil, wie Zeitzone und UTC-Offset, können zur Disambiguierung bei der Ortsauswahl verwendet werden. Mit diesen ist es realisierbar, den Einfluss der Ausreißer der Prognose zu halbieren.
Abhängig von der gewählten Option verbessert sich die Genauigkeit und zugleich verringert sich auch die Anzahl (Recall) der verortbaren Tweets. Das Ergebnis des Ansatzes lässt sich folgendermaßen formulieren: Die Anzahl und die Genauigkeit verhalten sich konträr zueinander. Es können 56,77 % der Tweets mit einem Median von 15,22 km geortet werden und im Vergleich dazu nur 6,02 % mit einem Median von 3,39 km. Der Anwender kann die Genauigkeit der Lokalisierung im gegebenen Rahmen somit wählen. Das entwickelte Verfahren kann somit überall dort eingesetzt werden, wo keine exakte Verortung benötigt wird und der Nutzer sich in der Nähe des eingetragenen Standorts befindet. Mögliche Einsatzszenarien sind somit bei sozialen Erdbebensensoren, bei der Meinungsforschung und bei der Verfolgung der Ausbreitung von Seuchen. Weitere Szenarien sind Katastrophen bei Veranstaltungen mit begrenztem Einzugsgebiet. Im Vergleich zu den anderen Verortungsmethoden weist das in dieser Arbeit vorgestellte Verfahren nur eine geringe Rechenkomplexität auf, ist weltweit einsetzbar und schränkt keine Tweeter in der Nachrichten- oder Beziehungsanzahl ein. Es wird nur das Standortfeld aus dem Nutzerprofil des Tweeters benötigt.

Inhaltsverzeichnis

1. Einleitung

In diesem Abschnitt wird die Motivation für diese Arbeit verdeutlicht. Es werden die Problemstellung und die Zielsetzung erläutert. Anschließend werden die Bedeutung und die Relevanz der Sozialen Medien für den Katastrophenschutz erklärt.

1.1. Motivation

Im Katastrophenfall sind die Entscheidungsträger auf den Zugriff einer möglichst umfassenden Informationsbasis angewiesen, um die Hilfs- und Gefahrenabwehrmaßnahmen zu koordinieren. Soziale Medien, wie Twitter, sind weit verbreitet und Nutzer setzen diese ein, um Nachrichten in Krisenzeiten in Echtzeit auszutauschen. Im Katastrophenschutz werden die Sozialen Medien aktuell eher selten als Informationsquelle durch Behörden verwendet. Dabei enthalten sie krisenrelevante Fakten, welche normalerweise nicht über Notrufzentralen zugänglich sind, und können als Katastrophenindikatoren fungieren, um Gefahrensituationen vorherzusagen. Die Vision dieser Arbeit lautet: Die zukünftigen IT-Systeme der Entscheidungsträger können wichtige Informationen aus den Sozialen Medien verarbeiten, um eine bessere Entscheidungsfindung und eine präventive Schadensabwehr zu ermöglichen.

Absatzmittler und Produzenten von Konsumgütern sind heute besonders an persönlich zugeschnittener Werbung interessiert. Je besser die Werbung auf den Kunden abgestimmt ist, umso höher ist die Chance, dass die beworbenen Produkte abgesetzt werden. Neben der inhaltlichen Abstimmung hat der lokale Fokus an Bedeutung zugenommen. Mit der Lokalisierung des Kunden ist es möglich, ihm Produkte und Dienstleistungen in seiner nächsten Umgebung anzubieten.

Meinungsforschungsinstitute und politische Parteien haben ein essentielles Interesse an aktuellen Umfragen und Stimmungen. Mit Twitter ist es möglich, automatisierte Meinungsforschung durchzuführen. Je genauer die Stimmung geografisch lokalisiert werden kann, umso genauer können regionale Unterschiede analysiert werden. Maynard und Funk (1) entwickelten ein System zur automatischen Entdeckung von Meinungen in Tweets. Ihr System klassifiziert die politische Meinung zu 62,2 % korrekt (precision).

1.2. Problemstellung

Soziale Medien, wie Twitter, finden aktuell als Informationsquelle im Katastrophenschutz nur bedingt Verwendung. Eine Ursache hierfür ist die fehlende Verortung der Nachrichten. Die in den Nachrichten enthaltenen Informationen können keiner Schadenslage zugeordnet werden, da aktuell nur 0,86 % der Tweets mit einem geografischen Bezug versehen[1] sind.

In dieser Arbeit soll eine Methode entwickelt und evaluiert werden, welche die geografische Position des Nutzers zur Zeit des Absendens einer Twitternachricht bestimmen kann. Hierzu soll das im Nutzerprofil gesetzte Standortfeld verwendet werden. Der darin enthaltene Text kann neben einem Ort auch nicht geografische Informationen beinhalten. Zur eindeutigen Bestimmung von Ortsangaben ist es notwendig, aus diesem Text die darin enthaltenen Toponyme[2] zu extrahieren. Abbildung 1 zeigt das Standortfeld im Twitterprofil.

[1] Erläuterung in Kapitel 2.2.2. Übersicht über die Domäne Twitter.

[2] Definition Toponym: Bezeichner für geografische Objekte (91).

Bei der Toponymextraktion und Standortbestimmung stellen sich folgende Herausforderungen:

- Es muss untersucht werden, welche Informationen das Standortfeld beinhaltet und wie sie maschinell auszuwerten sind.
- Die ausgelesenen Toponyme müssen geografisch lokalisiert werden.

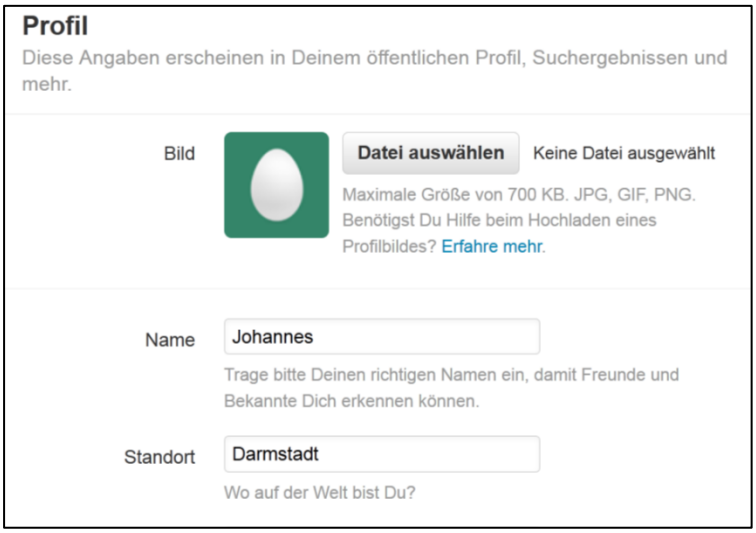


Abbildung 1: Nutzerprofilansicht von Twitter (22)

1.3. Zielsetzung der Arbeit

Das Ziel dieser Arbeit ist es, einen Ansatz zu entwickeln, um Twitternachrichten automatisiert zu verorten. Hierfür erfolgt eine Analyse der aktuellen Ansätze zur Lokalisierung von Tweets. Darauf aufbauend, soll eine Methode entwickelt werden, die anhand der Informationen aus dem Nutzerprofil Tweets verortet. Dabei wird der Standort als Approximation der Position des Tweets verwendet. Um die Güte der Prognose zu ermitteln, erfolgt die Entwicklung eines Modells zur Messung der Distanz von der Position des Tweets und dem Ort aus dem Standortfeld. Zur Evaluierung der Methode wird ein Prototyp implementiert.

1.4. Relevanz der Arbeit

Soziale Medien, wie Twitter, können im Krisenfall für den Katastrophenschutz und für die Betroffenen eine Bedeutung als Informationsquelle und Vorortreport einnehmen, wie Palen *et al.* (2) berichten. Sie erklären weiterhin, dass die Nutzung von Twitternachrichten zuvor einer Verortung bedarf. Die Informationen aus verorteten Tweets können beispielsweise auf Landkarten dargestellt werden. Wie bedeutsam Kartenmaterial im Krisenmanagement ist, erläutern Gunawan *et al.* (3). Plattformen, wie *Ushahidi* (4) und *Twitcident* (5), haben begonnen, krisenrelevante Tweets auf Karten darzustellen. Weiterhin berichtet Munro (6) anhand des Haitiunglücks, wie bedeutsam digitale Nachrichten[3] und ihre Lokalisierung im Krisenfall sind. Bezüglich des Beispiels von Haiti hatten 85 % der Bevölkerung die Möglichkeit, SMS zu versenden, und es trafen 40.000 Hilfegesuche über diesen Weg ein. Nur mit einer maschinellen Verortung ist es möglich, relevante Informationen aus der immensen Datenmenge in Echtzeit zu extrahieren. Um die Sozialen Medien im Katastrophenschutz effizient nutzen zu können, ist es somit notwendig, dass Tweets automatisiert lokalisiert werden. Diese Arbeit stellt einen Ansatz zur Lokalisierung vor.

[3] Munro (6) versteht unter digitalen Nachrichten vor allem SMS und Twitternachrichten.

1.5. Aufbau der Arbeit

Das Kapitel 2 erläutert die Grundlagen. Dazu gehört die Vorstellung von Twitter und der Lokalisierung von Tweets relevanten geografischen Themen. Im Kapitel 3 werden aus verwandten Arbeiten Methoden zur Tweetlokalisierung vorgestellt und miteinander verglichen. Kapitel 4 beschreibt die Umsetzung des Ansatzes mit der Vorstellung des Modells und der Erläuterung der Implementierung.
Das Kapitel 5 stellt die Ergebnisse vor und beschreibt die Evaluierung der einzelnen Optimierungen. Abschließend erfolgt in Kapitel 6 eine Zusammenfassung der Erkenntnisse und es wird ein Ausblick auf zukünftige Arbeiten formuliert.

2. Grundlagen

Das Kapitel Grundlagen führt zu Beginn in die Nutzung und Einsatzmöglichkeiten der Sozialen Medien im Katastrophenschutz ein. Anschließend erfolgen die Vorstellung und Analyse von Twitter. Mit einem Überblick über die wesentlichen geografischen Themen schließt das Kapitel.

2.1. Soziale Medien im Katastrophenschutz

In diesem Abschnitt werden zuerst die Verbreitung und Nutzung der Sozialen Medien im Krisenfall vorgestellt. Es werden die Einsatzmöglichkeiten von Twitter durch Behörden erläutert. Der darauf folgende Abschnitt widmet sich den existenten Lösungen zur Nutzung der Sozialen Medien im Katastrophenschutz.

2.1.1. Verbreitung der Sozialen Medien und deren Nutzung im Katastrophenfall

Das US-amerikanische Rote Kreuz untersuchte in einer Studie (7) den möglichen Einsatz von Sozialen Medien im Katastrophenfall. Befragt wurden 1058 repräsentativ ausgewählte US-Amerikaner in einer Online-Befragung im August 2010. Knapp 3 von 4 Befragten sind in einer Online Community[4]. Die am weitesten verbreiteten sind Facebook (58 %), Youtube (31 %), Myspace (24 %) und Twitter (15 %). Die Unterschiede in der Nutzung der Sozialen Medien korrelieren mit dem Alter und der Familiensituation der Befragten. Die Altersgruppe der 18 bis 34-Jährigen ist zu 89 % in einer Community vertreten, dem gegenüber nutzen die Befragten im Alter über 35 Jahre diese Dienste nur zu 65 %. Haushalte mit Kindern sind um 14 % häufiger vertreten.
In der Abbildung 2 ist die Häufigkeit des Zugriffs auf die Soziale Medien veranschaulicht, daraus wird ersichtlich, dass 50 % der befragten US-Amerikaner diese täglich nutzen.

Im Katastrophenfall würden sich aktuell 16 % der Befragten über Soziale Medien Informationen einholen und 18 % auf diesem Weg Notrufe absetzen, wenn die Notrufzentrale nicht erreichbar ist. Am populärsten zum Mitteilen von Augenzeugenberichten sind Facebook, Blogs und Twitter.

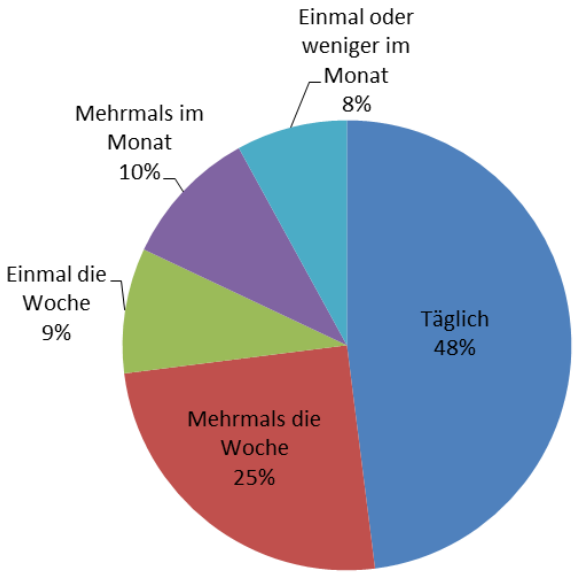


Abbildung 2: Häufigkeit der Nutzung der Sozialen Medien vgl. ARK (7)

[4] Community engl. für Gemeinschaft.

2.1.2. Einsatzmöglichkeiten von Twitter im Katastrophenschutz

Loveparade 2010 in Duisburg

Die Loveparade ist eine international bekannte Technoparade und fand im Jahr 2010 in Duisburg statt. Dabei ereignete sich am 24. Juli 2010 ein schweres Unglück mit 21 Todesopfern und 500 Verletzten (8). Aufgrund der Überfüllung an einem Tunnel kam es zu einer Massenpanik. Die Gäste der Veranstaltung twitterten bereits eine Stunde vor dem Unglück Nachrichten über das dichte Gedränge auf dem Festgelände (9). Der Nutzer *Sektorkind* schreibt: *„Hier ist echt kein Durchkommen. Ich versuche es trotzdem :)“*. Während und nach der Panik wurden Hilfegesuche, Informationen und Fotos ausgetauscht. In einer derartigen Lage ist es von Vorteil, Meldungen und Stimmungen direkt von Menschen, die vor Ort sind, zu erhalten. Eine Stimmungsveränderung kann als ein Indikator genutzt werden, um eine Massenpanik zu erkennen und damit präventiv agieren zu können. Mit Twitter besteht zudem die Möglichkeit, im direkten Kontakt die Betroffenen zu informieren.

Pukkelpop 2011[5]

Am realen Beispiel des belgischen Rockfestivals Pukkelpop vom 18. August 2011 wird deutlich, dass Twitter für die schnelle Informationsgewinnung und für die Organisierung von Hilfe bei großen Schadenslagen geeignet ist. Um 18:15 Uhr überraschte ein Unwetter das Festival und mehrere Bühnen und Zelte stürzten sturmbedingt zusammen. Die Folge waren zahlreiche Verletzte und fünf Todesopfer (10). Schon 15 Minuten zuvor kündigten Tweeter den Sturm in den Tweets mit *„You ready Pukkelpop?! Might be a wet one out there! Slip -n-slide-an-dog-an-you!! Remember that one @jaredle”* sowie mit *„I hope the storm does not burst on the Pukkelpop pasture. #PP11”* an.
Im Moment der Katastrophe ist es notwendig, alle Details schnell zu erfassen, um eine optimale Koordination der Hilfe zu gewährleisten. Während des Sturms berichteten die Festivalgäste über auftretende Schäden. Der Tweeter *TommyPorte* schrieb um 18:25 Uhr: *„Heavy storm at pukkelpop. Chateau tent is blown away. There is panic. I am sheltering in a toilet. #pp11“*. Die Abbildung 3 zeigt das erste Foto eines einstürzenden Zelts. Mit diesen krisenrelevanten Informationen erhält der Krisenstab einen wesentlich schnelleren Überblick über den aktuellen Zustand.
Nach dem Unwetter verbreiteten sich Nachrichten wie: *„Do you want to help? Use #hasselthelpt“* und *„#hasselthelpt […] If someone needs a place to sleep, let me know. We are from Hasselt #pp11”*. So organisierten die Gäste des Festivals und naheliegende Anwohner selbstständig Hilfe und Übernachtungsmöglichkeiten.

Die Abbildung 3 zeigt die Anzahl der Tweets, welche dem Pukkelpop zuzurechnen sind. Es ist ein deutlicher Anstieg der Nutzung während des Sturms und nach dem Sturm zu erkennen. Um 21:01 war der Höchstwert von 576 Tweets pro Minute erreicht.

Abbildung 3: Linkes Bild: Erstes Foto des einstürzenden Zelts „Chateau“ (87)
Rechtes Bild: Tweets pro Minute während des Pukkelpopfestivals nach Terpstra *et al.* (87)

[5] Die Twitternachrichten stammen von Terpstra *et al.* (87).

Straßenkrawalle in London

Ein weiteres reales Beispiel, bei dem die Sozialen Medien im Ernstfall nicht zur Unterstützung der Einsatzkräfte verwendet wurden, waren die Straßenkrawalle am 9. August 2011 in London. Gewalt bereite Jugendbanden organisierten sich zu Plünderungen und Vandalismusaktionen über den Blackberry Messenger und Twitter.
So rief der Twitternutzer *DanielNothing* zur Teilnahme am Aufstand auf: *„Heading to Tottenham to join the riot! who´s with me? #ANARCHY“* (11). Für die Entscheidungsträger wäre es relevant gewesen, zu wissen, an welchen Orten es zu den Ausschreitungen kommen würde. Stattdessen verfügte die Londoner Polizei nur über geringe Kenntnisse bezüglich der aktuellen Brennpunkte der Randalierer. Die Jugendlichen aber nutzten die neuen Medien, um sich über die Ankunft der Polizei untereinander zu informieren. Dadurch gelang es ihnen, den Ordnungshütern geschickt auszuweichen und sich gleichzeitig an neuen Treffpunkten zu verabreden (12).
Die Bilanz der Ausschreitungen lautete: Fünf Todesopfer und beträchtliche Sachschäden (13).

Ausbreitung von Krankheiten

Zur Eindämmung von ansteckenden Krankheiten ist es notwendig, ihre Ausbreitung genau zu verfolgen. Twitternachrichten können relevante Informationen über die Verbreitung beinhalten und somit einen Mehrwert für die Seuchenbekämpfung generieren. Mit dem *„Flu Detector“* System (14) (15) ist es möglich, zum Beispiel die Verbreitung der Grippe zu verfolgen. Lampos *et al.* (16) wiesen eine Korrelation von 95 % zwischen den Grippedaten aus Twitter und den offiziellen Angaben des Gesundheitsministeriums nach.

Die Geschwindigkeit bei der Katastrophenerkennung mit Twitter

In Notfällen spielt die Informationsgeschwindigkeit eine entscheidende Rolle. Wie schnell die Entdeckung von Unfällen mit Twitter vor sich gehen kann, wurde besonders deutlich am US-Airways Flug-1549. Ein Airbus A320 musste aufgrund eines Vogelschlags im Hudson River notwassern. Vier Minuten nach dem Absturz twitterte ein Beobachter: "*I just watched a plane crash into the hudson riv in manhattan*" (17) und kurz darauf wurde ein erstes Foto von einem Ersthelfer getwittert (Abbildung 5). Der Text zum ersten Foto lautete: *“There's a plane in the Hudson. I'm on the ferry going to pick up the people. Crazy.”* von Janis Krum.
Die Tweeter entdeckten das Unglück 15 Minuten schneller und lieferten detailliertere Informationen als die traditionellen Medien.

Abbildung 4: Twitterfoto des notgewasserten Flugzeugs im Hudson River (82)

2.1.3. Bestehende Lösungen zur Nutzung der Sozialen Medien im Katastrophenschutz

McClendon und Robinson (18) analysierten den Einsatz der Sozialen Medien als Informationsquelle im Katastrophenschutz. Dabei untersuchten sie speziell *Ushahidi* (4) und *Tweak the Tweet* (5).
Ushahidi ist eine Open Source Plattform, welche zur Informationsauswertung in Katastrophenfällen dient. Dabei werden verschiedene Informationsquellen, wie beispielsweise SMS und Twitter, verarbeitet und auf Karten angezeigt. Zur Bearbeitung, Klassifizierung und Verortung der Daten kommen freiwillige Helfer (eine Crowd) zum Einsatz. Die durch die Plattform gewonnenen Informationen unterstützten die Mitarbeiter des Katastrophenschutzes bereits in zahlreichen Krisenfällen, besonders bekannte Einsätze waren die Erdbeben 2010 in Haiti und Neuseeland.
Tweak the Tweet ist eine speziell für Katastrophen entwickelte Twittersprache und ein Modul für Ushahidi. Dabei werden in den Nachrichten Hashtags, wie #location, #status, #needs und #damage, verwendet, um die Informationen besser zu verarbeiten. Tweets mit dieser Syntax sind im Twitterstream besser auffindbar.
Die *SwiftRiver* (19) Plattform stellt Module zur Filterung, Analyse und Georeferenzierung für Ushahidi zur Verfügung, laut McClendon und Robinson (18).

Tweetincident (20) ist ein Web-basiertes Framework, welches die Analyse, Filterung und Suche nach Störfällen oder Krisen aus sozialen Webstreams ermöglicht, Abel *et al.* (21). Abbildung 6 verdeutlicht, wie Tweetincident funktioniert. Zuerst erfolgt die Aufnahme von Störfällen über einen Broadcast der Behörden [1]. Dieser publiziert automatisch alle gemeldeten Krisen mit Beschreibung und Ortsangabe [a]. Zu den bekannten Fällen werden alle relevanten Tweets automatisiert gesucht [2, 3] und anschließend angezeigt [4].

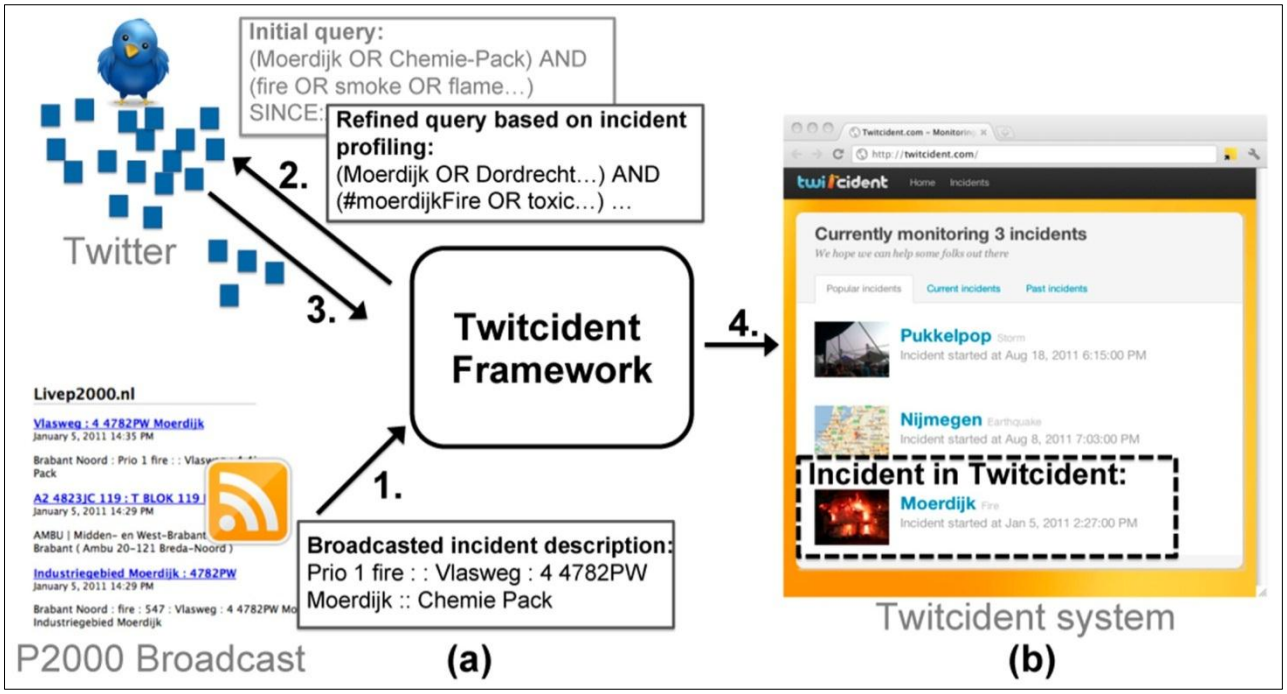


Abbildung 5: Funktionsweise von *Tweetincident* (20) - Abel *et al.* (21)

[1] Broadcast publiziert Unfall;
[2,3] relevante Tweets werden gesucht;
[4] Präsentation der krisenrelevanten Informationen;
[a] Ablauf des Systems ;
[b] Ergebnisansicht im Browser

2.2. Vorstellung und Analyse von Twitter als Mikrobloggingdienst

Die folgenden drei Abschnitte stellen den Mikrobloggingdienst Twitter vor, erläutern die Domäne und analysieren die Struktur der Twitternachrichten.

2.2.1. Twitter und Mikroblogging

Twitter (22) ist eine schnell wachsende Onlineplattform für Mikroblogging. Unter Mikroblogging versteht man das Kommunizieren von kurzen, SMS-ähnlichen Textnachrichten, die chronologisch als Blog dargestellt sind. Twitternachrichten, auch Tweets genannt, ermöglichen es den Autoren, öffentlich über ihren persönlichen Status oder beliebige Themen zu berichten[6]. Tweets sind 140 Zeichen lang und werden in Echtzeit an Abonnenten (sog. *Follower*) versendet. Dabei wird das Absetzen einer Statusnachricht umgangssprachlich als „Twittern" bezeichnet. Eine spezielle Syntax erlaubt es, mit dem *@Benutzername* den Tweet an eine bestimmte Personen zu adressieren und mit dem Hashtag *#Thema* die Nachricht einer Thematik zuzuordnen. Mit dem Hashtag ist es möglich, Nachrichten zu einem Thema zu finden.

Twitter ist im März 2006 gegründet worden und hat aktuell über 465 Millionen Accounts (23) und 100 Millionen aktive Nutzer[7] (Stand: September 2011). Aktuell versenden die Tweeter über 200 Millionen Tweets am Tag (24). Getwittert wird über die Webseite und mithilfe mobiler Anwendungen. In der Abbildung 6 werden die acht am häufigsten vertretenen Nationen präsentiert und die Länder mit dem prozentualen Twittergebrauch unter den Internetnutzern. Die USA sind mit 107,7 Millionen Accounts am häufigsten vertreten. Aus den Ländern Niederlande, Türkei und Japan sind knapp über 30 % aller Internetnutzer bei Twitter angemeldet.

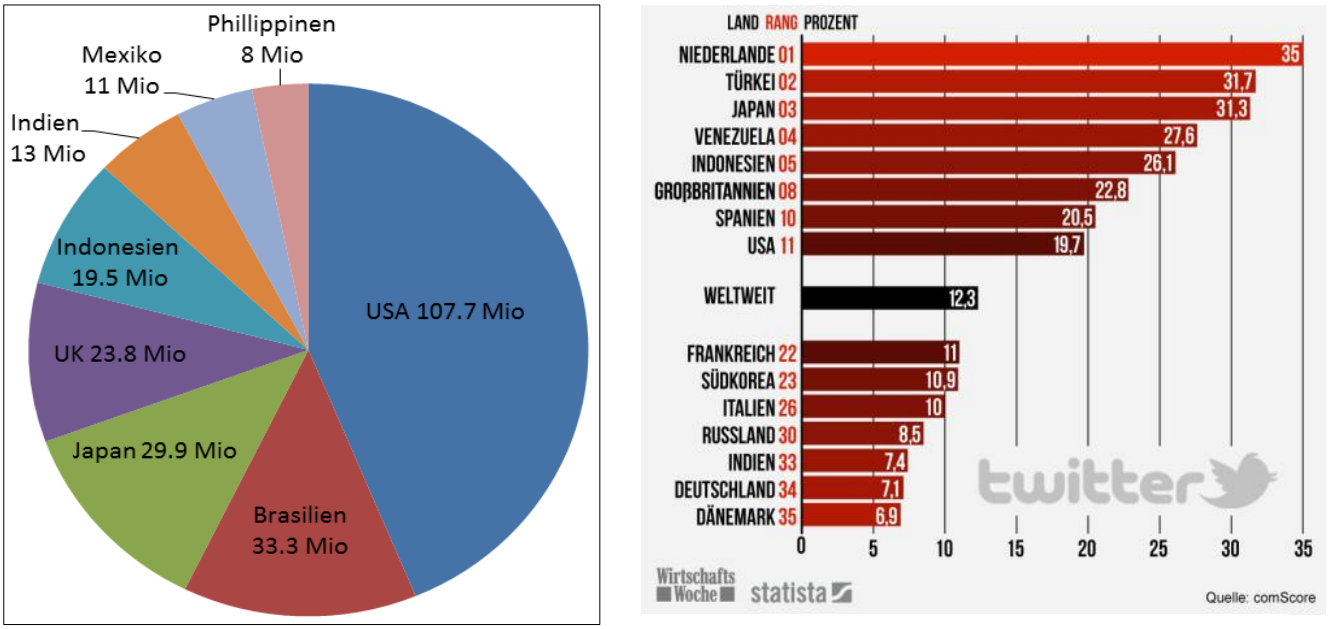


Abbildung 6: Links: Die acht Nationen mit den meisten Twitternutzern (23)
Rechts: Länder mit dem prozentualen Twittergebrauch unter den Internetnutzern (85)

[6] Mikroblogging vgl. Lohmann *et al.* (90).

[7] Definition *aktiver Nutzer:* Tweeter loggt sich einmal pro Monat ein (86).

2.2.2. Übersicht über die Domäne Twitter

Twitter ermöglicht es, über zwei Programmierschnittstellen (API) Zugriff auf die Daten von Nutzern zu erhalten. Mit der Streaming API erhält der Nutzer einen kontinuierlich fließenden Anteil an aktuell gewitterten Nachrichten. Die Funktionen zur spezifischen Suche nach Tweets oder Nutzern werden von der Search API zur Verfügung gestellt. Über diese Programmierschnittstellen erhält der Nutzer neben den im Profil ersichtlichen Angaben weitere Metadaten, wie Zeitzone und verwendete Sprache. In der Abbildung 7 wird eine Übersicht über die bereitgestellten Daten vermittelt.

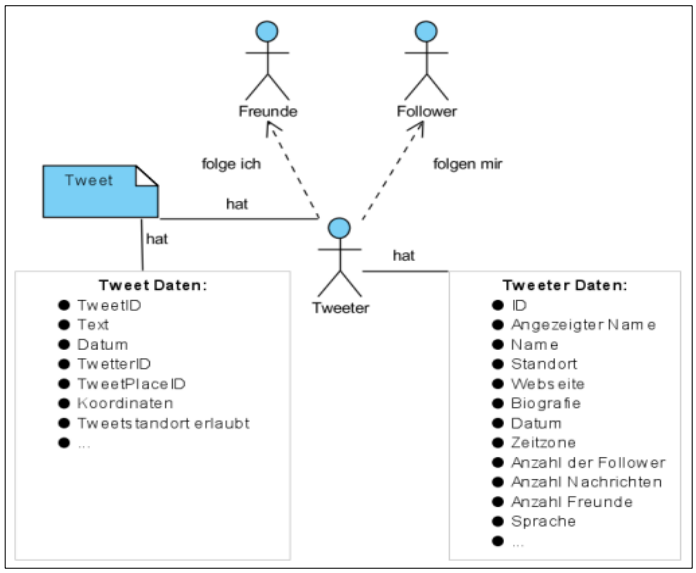


Abbildung 7: Verfügbare Informationen vom Tweeter (22)

Twitternachrichten können nicht nur ausschließlich über die Twitterwebseite versendet werden, sondern auch über mobile Anwendungen (Apps) und Webseiten von Drittanbietern (Tabelle 1). Die fünf meist verbreiteten Dienste wurden von Kinsella *et al.* (25) evaluiert. Hale *et al.* (26) weisen regionale Unterschiede in der Verbreitung dieser Dienste nach.

Dienst	Anteil der Tweets
Twitterwebseite	24,2 %
Foursquare	18,3 %
iPhone App	12,3 %
Android App	6,3 %
Ecofon	5,7 %

Tabelle 1: Die fünf häufigsten Quellen von Tweets nach Kinsella *et al.* (26)

Im August 2009 führte Twitter erstmalig die Möglichkeit ein, Tweets mit einer geografischen Position zu versenden (27). Tweets können dabei mit einer GPS-Position oder einem frei wählbaren geografischen Label, hier als *Place* bezeichnet, versehen werden.
Der Place kann verschiedene geografische Größenordnungen annehmen. In der Tabelle 2 erfolgt die Erläuterung der Placetypen. Dabei ist *Neighborhood* ein Polygon von mehreren Koordinaten. Mit diesem ist es möglich, geografische Gebiete exakt abzubilden. Die Abbildung 8 verdeutlicht die Umsetzung des Neighborhood Placetyps.

Takahashi *et al.* (28), Watanabe *et al.* (29), Kinsella *et al.* (25) und weitere untersuchten, wie häufig Tweets mit einer geografischen Position versendet werden. Nur 0,86 % aller Nachrichten haben einen geografischen Bezug, insgesamt 0,61 % in Form eines Place Labels. Ein Tweet kann ein GPS-Koordinatenpaar besitzen und zugleich ein Place Label. Dabei kann dieses Label redundant dasselbe enthalten. Insgesamt sind 0,54 % mit einer GPS-Position versehen, laut Kinsella *et al.* (25). Eigene Analysen ermittelten einen Anteil von GPS-kodierten Twitternachrichten in Höhe von 0,7 %. Es lässt sich subsumieren, dass insgesamt nur sehr wenige Tweeter ihre Nachrichten mit einem geografischen Bezug versenden.

Placetyp	**Erklärung**
Point	Eine GPS-Koordinate
Neighborhood	Besteht aus einem Polygon
City	Bereich einer Stadt
Admin	Verwaltungsbezirk, Bundesstaat
Country	Land

Tabelle 2: Erläuterungen der Placetyps (22)

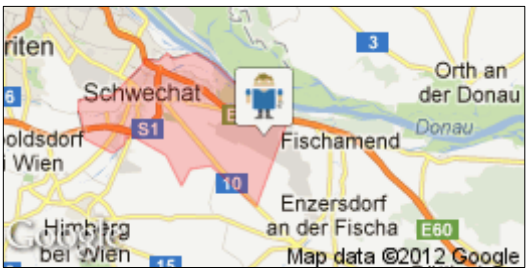


Abbildung 8: Darstellung des *Neighborhood* Placetyps (22)

2.2.3. Analyse der Struktur der Twitternachrichten

Analysen der Nutzung von Twitter und des Verhaltens der Tweeter wurden von Boyd *et al.* (30) durchgeführt. Der Schwerpunkt ihrer Arbeiten liegt auf dem *„Retweet“* [8] -verhalten und der statistischen Untersuchung der Nachrichteninhalte. Mit dem Wissen über den Inhalt der Tweets ist es möglich, alternative Lokalisierungsmethoden zu entwickeln.

Die Ergebnisse der Textanalyse von Boyd *et al.* (30) und von Kinsella *et al.* (25) lauten:

- 9,7 ± 6.8 Wörter enthält eine Twitternachricht,
  - das sind durchschnittlich 70,6 ± 40,4 Zeichen
- 36 % aller Tweets erwähnen einen Nutzer *„@Nutzer“*
  - 86 % beginnen mit *„@Nutzer“*
- 22 % aller Tweets beinhalten eine URL[9]
- 5 % aller Tweets enthalten einen Hashtag (#)
  - diese enthalten zu 41 % eine URL
- 3 % aller Tweets sind Retweets
  - Retweet-Kennzeichnungstyp: „RT“ 88%, 11% „via“ und „retweet“ mit 5%

Kinsella *et al.* (25) untersuchten eine kleine Twitternachrichtenmenge aus dem Jahr 2010, welche keine Duplikate enthielt. Im Vergleich zur Analyse von Boyd *et al.* (30) aus dem Jahr 2009 stellten sie eine Steigerung der Verwendung von Hashtags auf 11 % und bei den Usernamen auf 52 % fest.

[8] Ein Retweet ist die Antwort auf eine Twitternachricht.

[9] URL – Abk. für engl. Uniform Resource Locator - Zeichenfolge, die zur Lokalisierung einer Ressource dient. Vgl. Brockhaus (93) Band 28, S. 447.

2.3. Herausforderungen und Besonderheiten bei der Georeferenzierung

Dieser Abschnitt betrachtet die bei der Lokalisierung von Orten auftretenden Herausforderungen und Besonderheiten.

Eine besondere Herausforderung ist die geografische Ambiguität. Unter dem Begriff Ambiguität ist die Mehrdeutigkeit von Zeichen und Zeichenketten zu verstehen. Im geografischen Sinn bedeutet dies, dass ein Ortsbezeichner (z.B. Darmstadt) nicht eindeutig ist und mehrere reale Objekte[10] identifiziert. Mehrdeutigkeit in Bezug auf geografische Orte lässt sich in zwei Bereiche differenzieren:

1.) Ortsbezeichner benennen unterschiedliche Orte (**Geo/Geo**).
2.) Ortsbezeichner benennen neben dem Ort auch ein Ding oder eine Person (**Geo/Non-Geo**).

In den Bereich der **Geo/Geo**-Ambiguität lassen sich Ortsbezeichner einordnen, welche mehrere reale Orte bezeichnen. Beispielsweise existieren unter dem Ortsnamen *„Darmstadt“* eine Stadt in Deutschland und zwei weitere in den Vereinigten Staaten. Amitay *et al.* (31) haben ermittelt, dass in Webseiten 37 % der enthaltenen Ortsnamen eine mehrdeutige geografische Bedeutung haben. Nach Lieberman *et al.* (32) existieren neben der französischen Stadt Paris weitere 59 gleichnamige Orte.

Bei der **Geo/Non-Geo** Ambiguität können Bezeichner neben der Benennung von Orten gleichzeitig Personen oder Objekte benennen. Beispielsweise bezeichnet *„Vienna“* eine Stadt in Österreich und dient gleichzeitig als weiblicher Vorname.
Unter *„Metro“* [11] versteht man eine Untergrundbahn, eine Großhandelskette, eine Zeitung, eine Software, eine Musikgruppe, einen weiblichen Vornamen und eine Stadt in Indonesien.

Smith *et al.* (33) analysierten in historischen Texten vorkommende Ortsbezeichner auf Mehrdeutigkeit und verzeichneten dabei kontinentale Unterschiede in ihrer Häufigkeit. In der Tabelle 3 wird ein Vergleich der Kontinente vorgenommen. Nord- & Zentralamerika verfügt gegenüber Europa über sehr viele Ortsnamen, welche auf verschiedene Orte verweisen. Die größere Mehrdeutigkeit lässt sich durch die Tatsache erklären, dass bei der Besiedlung Amerikas die Siedlungen sehr häufig nach europäischem Vorbild benannt wurden, nach Smith *et al.* (33).

Kontinent	**Orte mit verschiedenen Ortsnamen**	**Ortsnamen, welche auf verschiedene Orte verweisen**
Nord- & Zentralamerika	11,5 %	57,1 %
Ozeanien	6,9 %	29,2 %
Südamerika	11,6 %	25,0 %
Asien	32,8 %	20,3 %
Afrika	27,0 %	18,2 %
Europa	18,2 %	16,6 %

Tabelle 3: Ambiguität von Ortsnamen nach Smith *et al.* (33)

[10] Ein Objekt ist allgemein ein Gegenstand. Vgl. Knauers Lexikon (89) S. 470.
[11] Vgl. Armitay *et al.* (31) und Brockhaus (93) Band 18, S. 354.

Neben der allgemeinen geografischen Ambiguität beinhaltet die textbasierte geografische Referenzierung einige weitere Herausforderungen. Woodruff *et al.* (34) fassen diese folgendermaßen nach (Farrar & Lerud 1982 und Griffiths 1989) zusammen:

- **Ortsschreibweisen**
  Es existieren für einen Ort verschiedene Schreibweisen.
  Beispiele: *Schwartza/Schwarza*, *Saßnitz/Sassnitz*
- **Alternative Namen**
  Es existieren mehrere alternative Namen für einen Ort.
  Oftmals ist dies sprachlich bedingt, wie *Wien/Vienna* und *Köln/Cologne/Colonia*.
- **Verschiedene Orte mit ähnlicher Schreibweise**
  Beispiele: *Bärental, Bärenthal, Baerenthal*
- **Homonyme**[12]
  Beispiele: *Konstanz/Stadt am Bodensee*, *Konstanz/Beständigkeit*
- **Neologismen**
  Es existieren Wortneuschöpfungen für bekannte Orte.
  Beispiele: *Chemnitz/Karl-Marx-Stadt*, *Sankt Petersburg/Leningrad*, *Saigon/Ho-Chi-minh-Stadt*
- **Akronyme**
  Häufig werden Orte umgangssprachlich in Texten abgekürzt. Beispiele sind hier *da/DA* für Darmstadt und *F/FFM* für Frankfurt am Main. Diese Abkürzungen sind oftmals nur im Kontext verständlich.

Weiterhin werden politische Grenzveränderungen und Namensänderungen erläutert, welche im Bereich von Twitter nur eine untergeordnete Rolle spielen.

Eine weitere geografische Herausforderung ist die Repräsentation von Orten mit nur einem Koordinatenpaar. Hecht *et al.* (36; 37) erläutern, dass in Wikipediaartikeln, auf Flickerbildern und in geografischen Lexika nur ein Koordinatenpaar statt eines Polygons angegeben ist. Bei Distanzmessungen führt dies zu erheblichen Ungenauigkeiten. Die Größe der Abweichung wird am Beispiel des Wikipediaartikels *Russland* (35) mit den Koordinaten 59° N, 70° O deutlich. Diese liegt mittig in einem Land mit einer Ost-West-Ausdehnung von 9.000 km[13].

2.4. Georeferenzierung unter Verwendung von geografischen Lexika

Ein geografisches Lexikon listet Toponyme mit eindeutigen Koordinaten auf. Unter Toponymen sind alle Ortstypen, wie Staaten, Städte, Flüsse, Berge, Täler und Straßen, zu verstehen. Ein geografisches Lexikon enthält oftmals Zusatzinformationen, wie die Population, alternative Ortsbezeichner und Abkürzungen. Es folgt ein Überblick über die aktuell verfügbaren geografischen Lexika, im Speziellen wird auf das Lexikon Geonames eingegangen.

[12] Homonyme :„Wort, das mit einem andern gleich lautet, den gleichen Wortkörper hat (aber in der Bedeutung [und Herkunft] verschieden ist“ Definition nach Duden (77).

[13] Vgl. Brockhaus (93) Band 23, S. 551.

2.4.1. Übersicht über verfügbare Lexika

Geografisches Lexikon	Umfang	Verwendbarkeit	Nutzungsbeschränkung pro Tag
MetaCarta (26; 36)	weltweit	kommerziell	keine Einschränkung
Google Geocoding API (26; 37)	weltweit	kommerziell	Webservice: 2.500 kostenfreie Anfragen
Yahoo PlaceFinder (26; 38)	weltweit	kommerziell	Webservice: 50.000 kostenfreie Anfragen
Geonames mit Gisgraphyclient (39)	sämtliche Toponymtypen weltweit, Zugriff auf Wikipedia	frei verwendbar	Webservice: 30.000 kostenfreie Anfragen lokale Installation: ohne Limit
GNIS - USGS Geographic Names Information System (40)	nur USA	frei verwendbar	lokale Datei, keine Einschränkung
United Nations Statistics Division (41)	Länder und Kontinente	frei verwendbar	lokale Datei, keine Einschränkung
OpenStreetMap (42)	weltweit, Straßen und Adressdaten	frei verwendbar	Webservice, lokale Datei, keine Einschränkung
World Gazetteer (43)	Länder, Verwaltungs-bezirke, Städte	frei verwendbar	lokale Datei, keine Einschränkung
Getty Thesaurus of Geographic Names (44)	weltweit	kommerziell	Webservice, lokale Installation
ISO Standard Akronyme (45)	Länder Akronyme weltweit	frei verwendbar	lokale Datei, keine Einschränkung
Wikipedia (46)	weltweit – sehr präzise	frei verwendbar	Webservice, lokale Installation, keine Einschränkung

Tabelle 4: Übersicht über geografische Lexika

2.4.2. Detailbetrachtung von Geonames und Gisgraphy

Gisgraphy (39) ist ein Framework, welches den Zugriff auf die geografischen Lexika *Geonames* und *OpenStreetmap* ermöglicht. Dabei umfassen die zwei Lexika zusammen über 42 Millionen Einträge. Die Einträge beinhalten Orte mit Koordinaten und Zusatzinformationen. Es existieren über 100 verschiedene Ortstypen, wie Straßen, Häfen, Städte, Ozeane und Kontinente. In Abhängigkeit vom Typ sind unterschiedliche Informationen erhältlich. Der Zugriff auf die Daten erfolgt über einen lokalen oder vom Gisgraphybetreiber zur Verfügung gestellten Webservice.

Gisgraphy bietet zahlreiche Funktionen zum Durchsuchen der Daten an:

Funktion *Georeferenzierung*:

Eine Koordinate wird zu einer gegebenen Adresse in Form einer Straße, Stadt, Postleitzahl gefunden. Die Eingabe kann ebenfalls in einer Kombination der oben genannten Adressformen erfolgen. Voraussetzung für die Suche ist die Angabe des Landes, in welcher sich die Adresse befindet.

Funktion *Reverse Georeferenzierung / Straßen Service*

Mit dieser Funktion sucht der Nutzer nach Straßen per Name oder GPS-Position. Zusätzlich ist es möglich, die Suche durch eine Radiuseingabe um ein Koordinatenpaar zu optimieren.

Funktion *Volltextsuche*

Mit der Volltextsuche ist es möglich, nach beliebigen Inhalten oder bestimmten Ortstypen zu suchen. Dabei ist die Suche auf Länder, spezielle Ortstypen oder Sprachen einschränkbar.

Rückgabewerte
Bei der Verwendung der Gisgraphy Funktionen erhält der Nutzer eine Menge von Orten als Rückgabewert. Die Ergebnismenge lässt sich hinsichtlich der Anzahl und des Informationsumfangs zuvor festlegen.

Funktionalitäten des Geonames Webservices

Der Anbieter von Geonames stellt mehrere Webservices (47) und eine Programmierschnittstelle (API) für diesen zur Verfügung, um auf die geografischen Daten zugreifen zu können. In der Tabelle 5 wird ein kurzer Überblick über die wichtigsten Funktionen vermittelt und anschließend erfolgt eine genauere Erläuterung der Suchmöglichkeiten mit der Volltextsuche.

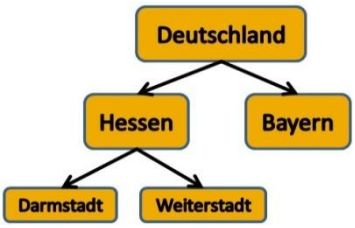
Name des Webservices	**Funktion**
GeoNames search	Stellt eine Volltextsuche zur Verfügung.
Placename lookup with postalcode	Findet zur Postleitzahl den passenden Ort.
Postal code country info	Enthält Ländernamen, Abkürzungen und Informationen zur Postleitzahl.
Find nearby: *toponym, postal codes, populated place, reverse geocoding*	Findet Orte/Toponyme im freiwählbaren Umkreis zu einem Koordinatenpaar oder einer Postleitzahl.
Place Hierarchy	Ermöglicht es, Toponyme in Bezug zu ihrer geografischen Hierarchie zu suchen. Ein Beispiel für eine solche Hierarchie bietet die Abbildung 9. So können Toponyme oberhalb, unterhalb sowie benachbart gefunden werden.  Abbildung 9: Toponymhierachie
Wikipedia	Findet Wikipedia-Artikel zu: Ortsnamen, GPS-Positionen, Postleitzahlen oder innerhalb eines Koordinatenpolygons.
Other	Webservices zu: Erdbeben, Wetter, Semanic Web, Verzeichnis von Ländern mit Koordinatenpolygon, Zeitzonen, GeonamesID-Suche

Tabelle 5: Vorstellungen der Geonames Webservices (47)

Volltextsuche

Bei der Volltextsuche (48) von Geonames stehen zahlreiche Optionen zur Verfügung, um die Suche einzuschränken, siehe Tabelle 6. Dabei kann nach der exakten Schreibweise des Suchbegriffs gesucht werden oder mit einer einstellbaren Ähnlichkeitssuche nach einem gleichartigen Ort.

Option	**Beschreibung**
Q	Sucht über alle Ortstypen.
Name	Nur Städte und Dörfer werden gefunden.
Name_equals (Name identisch)	Nur der Ort mit dem exakt übereinstimmenden Suchbegriff wird zurückgeliefert.
Feature Code (Ortstyp)	Beschränkt die Suche auf geografische Größenordnungen, analog zum Ortstyp bei Gisgraphy. So kann beispielsweise nur nach Häfen gesucht werden.
Name_startsWith (Ortsname beginnt mit)	Es ist erforderlich, dass der Ortsname mit der Zeichenkette der Suchanfrage übereinstimmt.
IsNameRequired (Ist ein Schalter, welcher aktiviert oder deaktiviert sein kann)	Mindestens ein Wort der Suchanfrage muss im Ortsnamen enthalten sein. Beispiel *„Berlin“*: Deaktiviert (false): findet alle Orte mit Berlin im Namen und die im Bundesstaat Berlin liegen.
Fuzzy	Ist ein Wert von Null bis Eins, der die Genauigkeit der Suche steuert. Je näher der Wert der Eins ist, desto genauer erfolgt die Suche.
Polygon mit vier Koordinaten (Ost, West, Nord, Süd)	Nur Orte innerhalb des Polygons werden als Ergebnis zurückgegeben.

Tabelle 6: Optionen bei der Volltextsuche (48)

Ausgabeformate von Geonames und Gisgraphy

Es werden die vier Ausgabeformate *Short*, *Medium*, *Long* und *Full* bereitgestellt. Das jeweilige Format beschreibt die Größe des Informationsumfangs und erweitert stets das kleinere Format.
Hier die Übersicht (Tabelle 7) über ausgewählte Inhalte der Ausgabeformate:

Short	**Medium**	**Long**	**Full**
Ortsname, Ortstyp, Land	Koordinaten, Straßenattribute, Zeitzonen, gesprochene Landessprachen	Informationen von übergeordneten Administrationsbezirken, alternative Namen	Alternative Namen der übergeordneten Administrationsbezirke

Tabelle 7: Übersicht über die Ausgabeformate (47)

2.5. Vorstellung der Textanalyseverfahren zur Unterstützung bei der Georeferenzierung

In diesem Abschnitt werden zwei Verfahren der Textanalyse angerissen. Diese finden häufig Anwendung bei der Toponymextraktion aus Texten.

Die *Named Entity Recognition* (kurz NER) ist ein Verfahren zur Extraktion von Eigennamen aus einem Text. NER Verfahren können beispielsweise aus einem Text Personen, Orte oder Zeiten extrahieren. Aktuelle Systeme identifizieren Eigennamen zu 93,39 % richtig und sind qualitativ fast gleichwertig gegenüber der manuellen Klassifikation mit 97,6 % (49), (50) und (51). Es existieren Verfahren der NER auf Basis der Sprache und auf Basis eines statistischen Modells, welches Trainingsdaten bedarf.

Part-of-speech Tagging (52) ist ein Verfahren aus der Computerlinguistik, welches die Wortart den einzelnen Wörtern eines Satzes zuordnet. Beispiel von Eugene Charniak (52): „Salespeople sold the dog biscuits“ wird zu Salespeople/**noun** sold/**verb** the/**det** dog/**noun** biscuits/**noun**.

3. Verwandte Arbeiten

In diesem Kapitel werden verwandte Arbeiten vorgestellt und miteinander verglichen. Zuerst erfolgt die Erläuterung der geografischen Themen, wie der Ortsangaben in Tweets, der Toponymextraktion und des geografischen Fokus von Texten. Anschließend folgen die Vorstellung und die Gegenüberstellung der bestehenden Ansätze zur Lokalisierung von Tweets.

3.1. Ortsangaben

Gerlernter und Mushegian (53) untersuchten die Struktur der Ortsangaben, welche in Twitternachrichten vorkommen können. Während der Untersuchung wurden 300 Tweets analysiert und die darin enthaltenen Orte nach Ortstypen klassifiziert. Angaben zur Häufigkeit der Orte in den Klassen liegen nicht vor.

Klasse	**Beispiele**
Staat, Bundesland, Region	Australien, Texas, Eichsfeld
Stadt, Stadtgebiet	New York, Berlin Mitte
Abkürzungen	AKL, CBD („Central Business District")
Infrastruktur	Lyttelton Port
Gruppierungen von Gebäuden	*Zeil* – Damit sind die Gebäude auf der Einkaufsstraße *Zeil* in Frankfurt gemeint.
Einzelne lokalisierbare Gebäude/Gebiete/Organisationen	Diamond Harbour School, Hutt Library
Was und Wo	BNZ in Riccarton, Shell Station in New Brighton
Straßen	75 Lyttelton Street
generische Orte	room, home, house, city
Orte mit HashTags	…. #Japan ready to go …

Tabelle 8: Aggregation der Ortsangabenanalyse nach Gerlernter und Mushegian (53)

3.2. Toponymextraktion und Ortsidentifikation

Zur Extraktion der Ortsangaben in einem Text und zur Identifikation des richtigen Orts werden hier die geeigneten Verfahren vorgestellt.

Verortung mit einem geografischen Lexikon

Eine sehr einfache Methode besteht darin, aus einem Satz die Wörter einzeln und in Phrasen in einem geografischen Lexikon zu suchen. Die Anzahl der Wörter pro einzelne Suchanfrage ist variabel. Beispielsweise kann im Satz *„Vienna ist sehr schön."* nach jedem Wort einzeln oder nach Kombinationen aus zwei oder drei Wörtern gesucht werden. Diese Methode identifiziert nicht, ob mit Vienna eine Person oder die bestimmte Stadt in Österreich gemeint ist.

Textanalyseverfahren

Mit einem Verfahren aus der Computerlinguistik, wie dem Part-of-Speech Tagging oder der Named-Entity-Recognition ist es möglich, Substantive zu erkennen. Aufgrund des Satzaufbaus kann eruiert werden, ob es sich um einen Ort oder eine Person handelt. Nur bei den Substantiven, welche als Orte identifiziert worden sind, ist es nötig, diese in einem geografischen Lexikon nachzugeschlagen. Ein Vergleich zwischen dem manuellen Auffinden von Toponymen in Tweets und dem maschinellen mit NER Software wurde von Gelernter *et al.* (53) vorgenommen. Dabei ist die NER Software bei der Toponymerkennung aus dem Text von Tweets zu 34,4 % bis 51,0 % erfolgreich, gegenüber 65,5 % bis 72,8 % bei der menschlichen Auswertung.
Es folgen zwei Beispiele der Toponymerkennung mit OpenCalais (54): *„Vienna/**City** is a nice city."*
*„My friend Vienna Müller/**Person** is buying a new Apple/**Company**."*
Eigene Experimente eruierten, dass OpenCalais nur solche Personennamen erkennt, welche in der Form *„Vorname Nachname"* angegeben sind.

GEO-System

Woodruff *et al.* (55) entwickelten ein *Geo-Referenced Information Processing System* (kurz *GIPSY*), welches geografische Namen in Texten automatisch den entsprechenden Koordinaten zuordnet (33). Das Verfahren filtert dabei ortsbeschreibende Worte, wie, „südlich von", „zwischen" und „nahe von" heraus, um eine genauere Positionsbestimmung zu ermöglichen.

Extraktion von Komma-Gruppen

Bei der Extraktion von Toponymen aus Texten spielen *Komma-Gruppen* eine besondere Rolle. Darunter sind die Phrasen „Ort, Land", wie beispielweise *„Darmstadt, Illinois", zu verstehen*. Diese Gruppen stellen oftmals eine präzisere Ortsbeschreibung dar als nur „Darmstadt".
Lieberman *et al.* (32) entwickelten zur Toponymerkennung in *Komma-Gruppen* eine spezielle Heuristik, welche zu 97 % korrekt identifiziert. Die Evaluation fand anhand von 87.000 englisch-sprachigen Zeitungsartikeln und Blogs statt.

Methode zur genaueren Identifikation von Orten

Eine Möglichkeit, sehr präzise Ergebnisse, aber mit geringem Recall zu erhalten, haben Hecht *et al.* (57), (58) vorgestellt. Bei diesem Verfahren kommt Wikipedia als geografisches Lexikon mit deaktivierter Ähnlichkeitssuche zum Einsatz. Es werden nur solche Orte gefunden, welche mit der exakten Zeichenkette der Suchanfrage übereinstimmen. Dabei werden die Groß- und Kleinschreibung nicht beachtet. Das Rückgabeergebnis ist das Koordinatenpaar aus dem Wikipediaartikel.

3.3. Geografische Fokusbestimmung von Texten

Unter dem geografischen Fokus eines Textes ist der Bezug zu einer geografischen Position zu verstehen. Die Frankfurter Allgemeine Zeitung schreibt beispielsweise einen Artikel über die Stadt New York. Hier richtet sich der Fokus auf die Stadt New York, jedoch die Herkunft des Artikels ist in Frankfurt zu verorten.
Für Suchmaschinen ist der geografische Bezug einer Webseite interessant. Dadurch besteht die Möglichkeit, beispielsweise Webseiten oder Texte über New York zu finden. Analog ist es bei Twitternachrichten bedeutsam zu wissen, ob ein Tweeter über einen Ort nur spricht oder sich tatsächlich dort befindet.
Amitay *et al.* (31) entwickelten ein System zur Fokusbestimmung von Texten, speziell für Webseiten. Dabei kann der Text den Fokus einer Stadt, eines Bundeslands, Lands oder Kontinents annehmen.
Zuerst muss zwischen der Herkunft der Webseite, also dem Erstellungsort und Hostingort, und dem inhaltlichen Fokus differenziert werden. Daten, wie die IP-Adresse, oder extrahierte Adressdaten aus dem Impressum sind der Herkunft zuzuordnen.

Ablaufschritte des Algorithmus zur Fokusbestimmung:

1. Text nach Toponymen scannen
Es werden nur Toponyme gefunden, welche in einem geografischen Lexikon vorkommen. Verwendet werden hier als Lexika GNIS (40), UNSD (41) sowie World-Gazetteer (43). Abkürzungen und häufige doppeldeutige geografische Terme werden herausgefiltert.

2. Disambiguierung[14]
Stehen hinter einem gefundenen Toponym ein weiteres oder eine Abkürzung, wie beispielweise „Darmstadt, Indiana“, so kann Darmstadt als Stadt in Indiana lokalisiert werden. Granulare Ortsangaben, wie eindeutige Straßen und konkrete Gebäude, werden als Stadt subsumiert. Bei mehreren gleichnamigen Toponymen wird die Stadt mit der größeren Population als die wahrscheinlichere angenommen. Bereits im Text gefundene eindeutige Ortsbezeichner werden bei der Identifikation eines einzelnen Toponyms verwendet. Nach der Disambiguierung liegt die folgende eindeutige Form *„Austin/Texas/United States/North America“* (speziell/generisch) mit einer Wahrscheinlichkeit von p vor. Die Wahrscheinlichkeit errechnet sich aus der Anzahl der im Text vorkommenden selben Orte, der Genauigkeit der Ortsangabe und der Ambiguität des Ortsnamen.

3. Fokusbestimmung des Textes
Der Algorithmus sortiert zuerst alle gefundenen Orte, absteigend nach ihrer Wahrscheinlichkeit p.
Anhand der Abbildung 13 ist ersichtlich, dass die Toponyme in einer Baumstruktur vorliegen und die unteren Blätter/Knoten in die Berechnung der oberen Knoten eingehen. Bei der Bestimmung der Fokusse wird die Baumstruktur genutzt, um eine Überdeckung zu erkennen. Mit Überdeckung ist die Situation gemeint, dass eine größere Region die kleinen Regionen überlagert. Damit wird eine Vergrößerung des Fokus verhindert. Beispielsweise überdecken die United States den Bundesstaat Texas und können somit verworfen werden. Bei Orlando in Florida und Texas liegt keine Überdeckung vor und somit wird Orlando als zweiter Fokus hinzugefügt. Irak und Asien werden aufgrund ihrer geringen Wahrscheinlichkeit verworfen.

[14] Disambiguation - Auflösung der Mehrdeutigkeit – siehe Kapitel 2.3.

6.41 Texas/United States/North America
4.97 United States/North America
4.2 Fort Worth/Texas/United States/North America
3.48 North America
1.68 Dallas/Texas/United States/North America
1.00 Orlando/Florida/United States/North America
0.70 Florida/United States/North America
0.56 Garland/Texas/United States/North America
0.25 Irak/Asia
0.17 Asia

Abbildung 13: Fokusalgorithmus von Amitay *et al.* (31)

Ergebnisse:
Der Fokusalgorithmus wurde von Amitay *et al.* (22) auf englischen Webseiten getestet. Er ordnet den Fokus mit 38 % dem exakten Ort, mit 68 % der Stadt oder dem Bundesland (engl. state) und mit 92 % dem Land richtig zu.

Smith *et al.* (16) entwickelten einen ähnlichen Algorithmus, welcher zusätzliche externe Informationen aus Biografien und Lexika verwendet und damit den geografischen Fokus von historischen Texten zu 74 % bis 93 % korrekt bestimmt. Die Tabelle 9 vergleicht die Ergebnisse des Algorithmus aus fünf verschiedenen Texten.

Text über	**Präzision (Precision) der Fokusbestimmung**	**Perfekte Disambiguierung**	**Recall der Erkennung von geografischen Informationen (Orte, Personennamen)**
Griechenland	93 %	98 %	99 %
Rom	91 %	99 %	100 %
London	86 %	92 %	96 %
Kalifornien	83 %	92 %	96 %
Mittleres Nordamerika	74 %	89 %	89 %

Tabelle 9: Ergebnisse des Fokusbestimmungsalgorithmus von Smith *et al.* (16)

3.4. Lokalisierung von Tweets auf Basis des Textes

Die Publikation von Paradesi (59) präsentiert ein System namens *TwitterTagger* zur Lokalisierung von Tweets auf Basis des Textinhalts. Dabei werden Substantivfolgen mit einem Part-of-Speech Tagger aus dem Text, beispielsweise *„I´m at **Holland Tunnel Toll Plaza**…"* (59), extrahiert.
Anschließend erfolgt ein Abgleich mit der USGS Geodatenbank (40).
Wird in der Datenbank kein passender Ort gefunden, so wird die Substantivfolge iterativ um das letzte Wort gekürzt. Es wird solange gekürzt, bis ein Eintrag gefunden wurde oder die Folge nur noch ein Substantiv besitzt und verworfen wird. Häufig in Tweets verwendete Wörter, wie „Love" und „Need", sind gleichzeitig Orte und finden somit keine weitere Beachtung.

Zur Auflösung der Mehrdeutigkeiten stehen zwei Module zur Verfügung, die Abbildung 10 stellt den Ablauf des Verfahrens dar.

Modul zur Auflösung der Geo/Non-Geo Disambiguierung

Dieses Modul verwirft alle gefundenen Substantivfolgen, welche keine ortsbestimmenden Präpositionen (in, nach, …) besitzen. Weiterhin wird geprüft, ob andere Tweeter dieselben ortsbestimmenden Präpositionen vor derselben Folge verwenden, anderenfalls wird diese verworfen. Damit ist sichergestellt, dass es sich um einen Ort handelt und nicht um eine Person. Das Beispiel von Paradesi (59) „She lives **in** Massachusetts" verdeutlicht die Intention zur Anwendung des Moduls. Es beinhaltet eine ortsbestimmende Präposition direkt vor dem Ort.

Modul zur Auflösung der Geo/Geo Disambiguierung

Es wird der Ort mit der kürzesten Distanz zwischen den gefundenen Orten und dem Standort[15] gewählt. Bei der Auswahl des Orts findet die Strecke zwischen den Standorten anderer Tweeter Verwendung, welche denselben Ort im Tweet erwähnt haben.

In der Tabelle 10 sind die Ergebnisse der Versuche von Paradesi angeben. Die Tweets können mit diesem System mit einer Precision von 15,81 % lokalisiert werden.

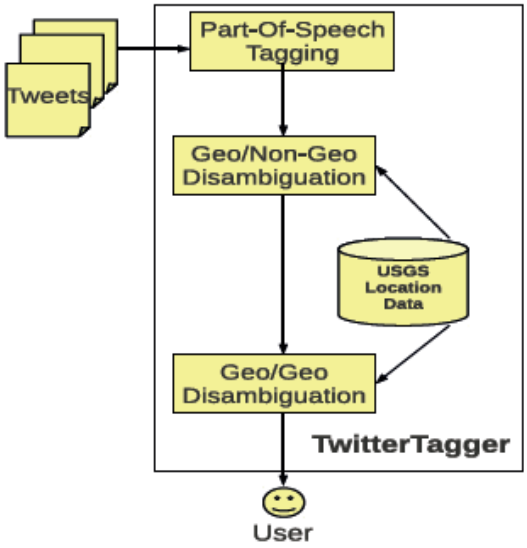


Abbildung 10: Twittertagger von Paradesi (59)

Metrik	Precision
Baseline	4,93 %
Geo/Non-Geo Disambiguierung	7,44 %
Geo/Non-Geo + Geo/Geo Disambiguierung	15,81 %

Tabelle 10: Ergebnisse des Twittertagers von Paradesi (59)

[15] Hier ist der Standort aus dem Standortfeld gemeint, Erläuterung im Kap. 1.2.

Verortung von Tweets anhand eines Sprachmodelles

Cheng *et al.* (27) verfolgten einen Ansatz auf ausschließlicher Basis des Texts. Die Lokalisierung auf der Stadtebene erfolgt nur anhand eines Sprachmodells, welches die regionalen sprachlichen Unterschiede der Twitternachrichten berücksichtigt. Einige Ausdrücke, wie das texanische *„howdy“* oder das hessische *„Gude“,* haben einen eindeutigen regionalen Bezug. Für jedes einzelne Wort wird mit der lokalen Häufigkeit ein statistisches Sprachmodell entwickelt. Nach einer Glättung und Filterung der Ausreißer kann einem Wort mit einer Wahrscheinlichkeit p ein Koordinatenpaar zugewiesen werden, siehe Abbildung 11, rechtes Diagramm.

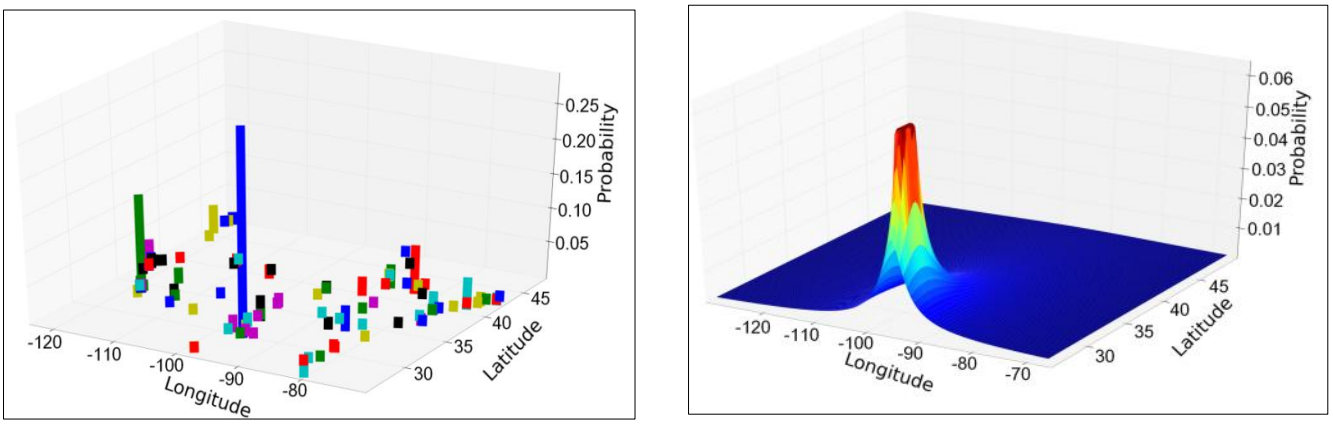


Abbildung 11: Links: Baseline des Sprachmodells für das Wort „rockets“ nach Cheng *et al.* (27)
Rechts: Sprachmodell für das Wort *„rockets“* nach der Glättung und Filterung, nach Cheng *et al.* (27)

Die Testdatenbasis besteht aus Tweetern, welche im Standortfeld Koordinaten eingetragen haben und über mindestens 1000 Nachrichten verfügen. Als Referenz (Baseline) wird das Sprachmodell ohne Filterung und Glättung verwendet. In dem linken Diagramm der Abbildung 11 ist dies beispielhaft für das Wort „rockets“ dargestellt.

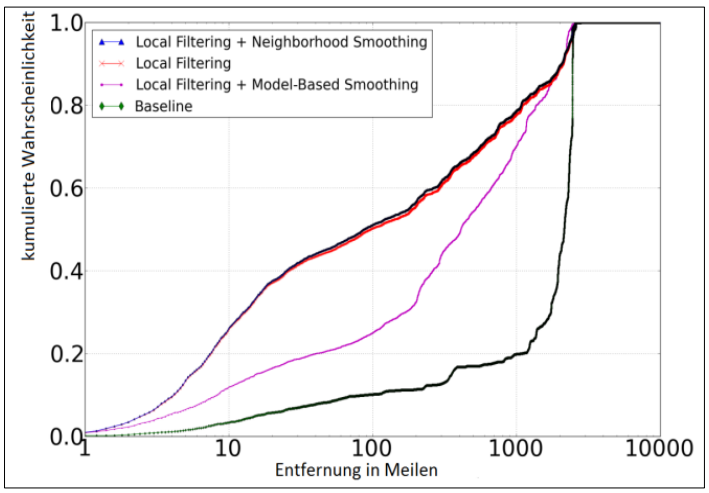


Abbildung 12: Verteilung der Entfernung in Bezug auf die Wahrscheinlichkeit den Tweet zu verorten nach Cheng *et al.* (27)

Mit dem Ansatz von Cheng *et al.* (27) können 30 % der Tweets mit einer Abweichung von bis zu 16 km und 50 % mit 161 km um ihre Position lokalisiert werden. Der Median entspricht somit in etwa 161 km. Die durchschnittliche Abweichung beträgt 862 km. Der Abbildung 12 ist die Verteilung der Entfernung zwischen dem prognostizierten Ort und dem realen Ort des Tweets zu entnehmen.

In dem Diagramm sind verschiedene Kurven von Glättungs- und Filterungsmöglichkeiten verzeichnet. Am besten schneidet eine Filterung aus lokalen Wörtern und einer Glättung auf die Größe eines Quadratgrads[16] (Neighborhood Smoothing) ab.

Den Zusammenhang zwischen der Anzahl der Tweets und der Größe des Mittelwerts erläutert das Diagramm 13. Je mehr Tweets vorliegen, umso genauer erfolgt die Lokalisierung.

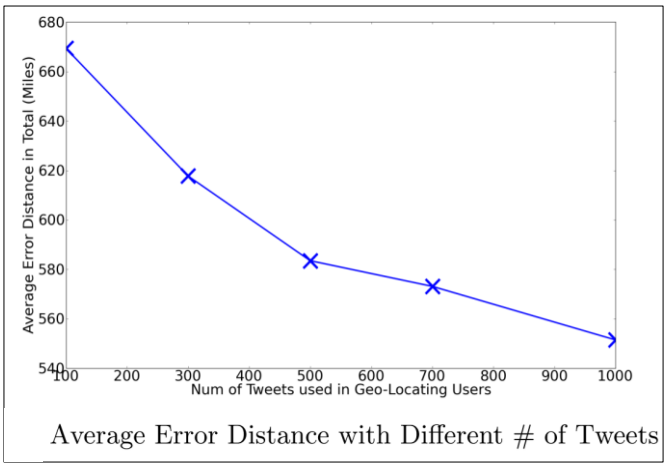


Abbildung 13: Mittelwert in Bezug auf die Anzahl der Tweets des Nutzers nach Cheng *et al.* (27)

Kinsella *et al.* (25) entwickelten ein Ortssprachenmodell, welches einem bestimmten Ort einen Korpus an Wörtern und Phrasen zuweist. Hierfür werden aus GPS-kodierten Tweets Ausdrücke extrahiert und dem Koordinatenpaar zugeordnet. Zum Beispiel wird zur Erstellung eines Ortssprachenmodells für Heidelberg der Textkorpus mit ortsspezifischen Wörtern, wie *Bismarckplatz*, *Neckar*, *Neckarwiesen*, *Ruprecht-Karls-Universität* und *Thingstätte gefüllt.*

Die Evaluierung verschiedener Prognoseverfahren ergab, dass das „Query Likelihood"-Verfahren am geeignetsten ist, um die Position zu bestimmen. Dabei wird für jeden Tweet die Position über das wahrscheinlichste Ortssprachenmodell eruiert. Tweets, welche über einen Ortsmitteilungsdienst (lokationsharing service), wie Foursquare, versendet wurden, sind mit diesem Modell nicht besser verortbar.

Genauigkeit auf Level von:	**Tweets korrekt verortet**	**Tweets gesendet über einen Ortsmitteilungsdienst**
Staat	53,2 %	51,4 %
Bundesstaat	31,6 %	24,6 %
Stadt	29,8 %	21,7 %
Postleitzahl	13,9 %	5,2 %

Tabelle 11: Ergebnisse des Modells von Kinsella *et al.* (25)

[16] Quadratgrad oder Raumwinkel genannt: Eine astronomische Einheit. Ungefähr 100 km². Vgl. Brockhaus (93) Band 22, S. 571.

Hecht *et al.* (58) stellten ein *Maschine Learning* -Experiment[17] vor, welches die Position des Nutzers anhand des Tweetinhalts lokalisiert. Dabei wird ein multinominales Naives Bayes Modell mit einem Vektor verwendet. Der Vektor besteht aus 10.000 Termen und der Häufigkeit jedes Terms.

Es wurden zwei Methoden der Vektorbildung getestet:

Count: - Zählt zu jedem in Tweets vorkommenden Wort dessen Häufigkeit und beinhaltet somit die 10.000 häufigsten Terme.

Calgari: - Ist eine Heuristik, um den Textcorpus mit überwiegend regionalen Wörtern anzureichern. Dabei ist die Annahme, dass weniger weitverbreitete und mehr lokale Wörter im Corpus die Klassifizierung verbessern. Häufig verwendete Wörter und Spam, wie „*lol*“, „*im*“ und „*love*“, werden dabei herausgefiltert.

Für die Erstellung von validen Trainingsdaten wurde der präzise Geotagger mit geringem Recall von Hecht *et al.* (57) (58) verwendet[18]. Ziel ist es, die Herkunft der Tweeter einem Land oder einem Bundesland zuzuordnen.

Es folgen zwei Strategien, die Stichproben für die Experimente auszuwählen.
Bei der *Uniform*-Strategie werden die Tweeter gleichmäßig aus den Ländern Großbritannien, Kanada, Australien und den USA ausgewählt.
Die Menge mit der *Random*-Strategie enthält zufällig gezogene Nutzer aus den vier genannten Ländern. Es ist zu beachten, dass Länder, wie die USA, eine größere Population haben und somit häufiger vertreten sind.
Für die Evaluation der Experimente wurde die Referenz (Baseline) an diese Größe der Stichprobe angepasst. Befinden sich beispielweise in der Menge Uniform 25 % Tweeter aus den USA, so beträgt die Referenz 25 %. Die Erklärung für den Referenzwert ist der Erwartungswert, da dieser bei einer zufälligen Wahl 25 % beträgt.

Experiment	**Anzahl**	**Methoden der Vektorbildung**	**Genauigkeit**	**Referenz (Baseline)**
Country-Uniform-2500	2500	Calgari	**72,71 %**	25 %
Country-Uniform-2500	2500	Count	68,44 %	25 %
Country-Random 20K	20.000	Calgari	**88,86 %**	82,08 %
Country-Random 20K	20.000	Count	72,78 %	82,08 %
State-Uniform-500	500	Calgari	**30,28 %**	5,56 %
State-Uniform-500	500	Count	20,15 %	5,56 %
State-Random 20K	20.000	Calgari	24,83 %	15,06 %
State-Random 20K	20.000	Count	**27,31 %**	15,06 %

Tabelle 12: Ergebnisse der Experimente von Hecht *et al.* (58)

[17] Vgl. Data Mining (94) S. 7.
[18] Siehe Kapitel 3.2.

Die Ergebnisse der Analyse der häufigsten Wörter mit lokalem Fokus sind in den Tabellen 13 und 14 aufgelistet. Zuvor wurden die Wörter auf den Wortstamm reduziert. Der „*Faktor*“ bedeutet, dass das Wort in dieser Region um den genannten Faktor häufiger verwendet wird als in anderen Regionen.

Wortstamm	**Land**	**Faktor**
„calgari“	Kanada	419,42
„brisban“	Australien	139,29
„coolcanuck“	Kanada	78,28
„clegg“	UK	35,49
„yelp“	USA	19,08

Tabelle 14: Die häufigsten Wörter mit lokalem Fokus auf Länderebene nach Kinsella *et al.* (25).

Wortstamm	**Bundesstaat (state)**	**Faktor**
„colorado“	Colorado	90,74
„elk“	Colorado	41,18
„redsox“	Massachusetts	39,24
„biggbi“	Michigan	24,26
„mccain“	Arizona	10,51

Tabelle 13: Die häufigsten Wörter mit lokalem Fokus auf Bundesstaatenebene nach Kinsella *et al.* (25).

Ikawa *et al.* (60) lokalisierten Tweets anhand der im Text vorkommenden Wörter. Zu jedem bekannten Ort in Ortsmitteilungsdiensten werden Tweets gesammelt und die dort enthaltenen Wörter extrahiert. Zuvor erfolgt eine Filterung der Retweets, da diese oftmals nicht aus demselben Ort stammen. Das Verfahren assoziiert zu jedem Ort einen Korpus von Wörtern. Die Tabelle 15 stellt zwei über das Format von Foursquare und Loctouch versendete Tweets dar.

Ortsmitteilungsdienst	**Nachrichten Format**
Foursquare	I`m at [Location] ([Address]) [Comment] (@[Location]).
Loctouch (in Japan)	[Location] にタッチ！[Comment] @[Location] にタッチ！

Tabelle 15: Nachrichtenformate der Ortsmitteilungsdienste Foursquare und Loctouch nach Ikawa *et al.* (63).

Verwendet ein Tweeter bestimmte Wörter aus dem Korpus, so kann ihm der dazugehörige Ort prognostiziert werden. Um die Qualität des Verfahrens zu vergleichen, ist die Referenz (Baseline) der Ort, an welcher der Nutzer die meisten Nachrichten in der Vergangenheit verschickt hat. Evaluiert wird das Verfahren mit der Korpusbildung für je einen einzelnen Nutzer und dessen Tweets und für alle Tweeter. Für alle Tweeter existiert nur ein Korpus pro Ort.
In der Abbildung 14 ist die Beziehung zwischen der Abweichungsdistanz und der Genauigkeit (precision) der Prognose in Bezug auf den tatsächlichen Ort veranschaulicht.

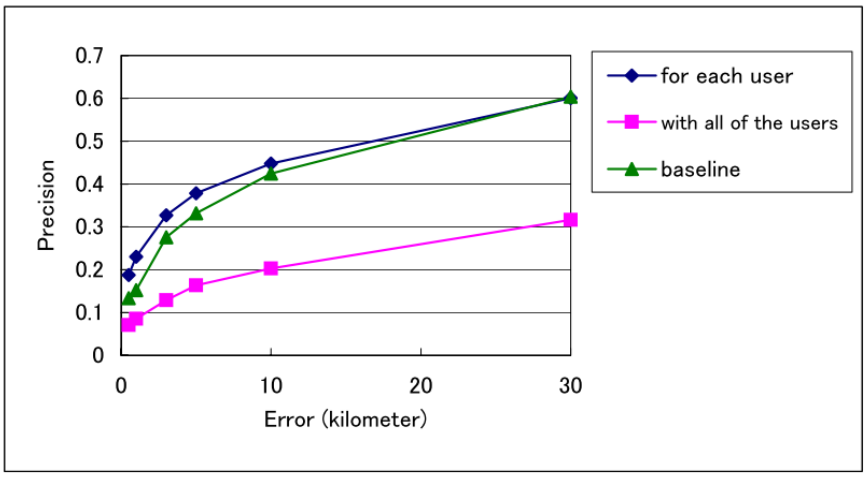


Abbildung 14: Zusammenhang Abweichung und Genauigkeit (precision) nach Ikawa *et al.* (60)

In der Tabelle 16 wird die Beziehung zwischen der Abweichungsdistanz und der relativen Häufigkeit der Verortung von Tweets verdeutlicht (Recall).

Radius in km	0,5	1	3	5	10	30
Für einzelne Tweeter	0,07	0,09	0,12	0,14	0,17	0,22
Für alle Tweeter	0,06	0,07	0,10	0,13	0,16	0,25

Tabelle 16: Beziehung zwischen der Abweichungsdistanz und Recall nach Ikawa *et al.* (60).

Eisenstein *et al.* (61) lokalisierten Tweets mithilfe der im Text vorkommenden regionalen Themen. Zur Bildung des geografischen Themenmodells wird das von Blei *et al.* (62) entwickelte *Latent Dirichlet Allocation* (LDA) Verfahren verwendet. Dabei wird zu jedem Thema ein geografischer Bezug hergestellt und so kann beispielsweise der Satz „*Lilien*[19] *gewinnen die Partie 4:0*“ mithilfe des Themenbereichs Fußball Darmstadt zu geordnet werden.
Als Datengrundlage kamen nur Tweets mit GPS-Koordinaten zur Verwendung, bei denen zu demselben Nutzer mindestens 20 Nachrichten vorliegen. Weiterhin wurden Tweeter herausgefiltert, welche mehr als 1000 Beziehungen haben. Analysen von Davis *et al.* (63) weisen nach, dass Tweeter mit zu vielen Kontakten ungeeignet sind, weil diese oftmals bekannte Persönlichkeiten oder Institute verkörpern. Die durchschnittliche Abweichung beträgt bei diesem Modell 900 km und der Median 494 km zwischen dem Tweet und der Prognose. Zu 24 % können Tweets einem US-Staat richtig zugeordnet werden. Analysen von Eisenstein *et al.* (61) ergaben, dass die Abweichung mit der Anzahl der geografischen Themen im Tweet zusammenhängt. In der Grafik 15 ist diese Beziehung dargestellt.

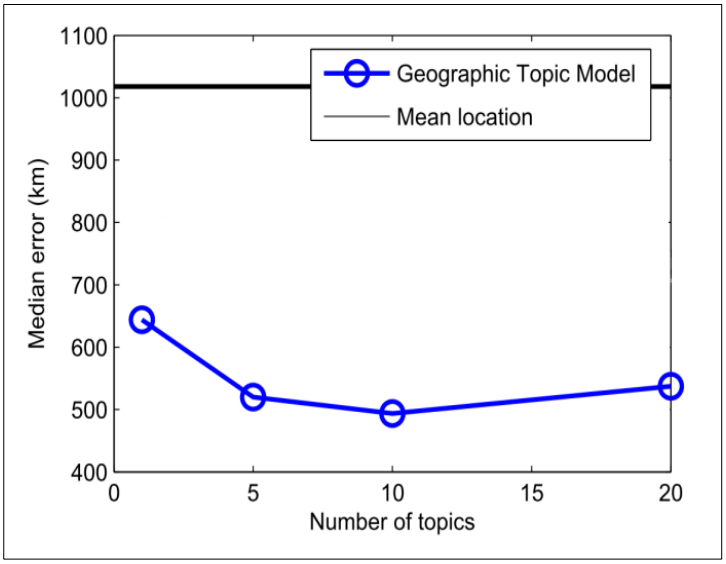


Abbildung 15: Vergleich Grafik von Eisenstein *et al.* (61).

[19] Der Sportverein Darmstadt 98 heißt umgangssprachlich „*Lilien*“.

Die Abbildung 16 präsentiert einige Beispiele von Orten mit den dazugehörigen regionalen Themen. Dabei bezeichnen die rot gedruckten Wörter fremdsprachliche Ausdrücke und die blau geschriebenen geben Eigennamen an.

	"basketball"	**"popular music"**	**"daily life"**	**"emoticons"**	**"chit chat"**
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
 Boston	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee	*ese* exam suttin sippin
 N. California	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	*pues* hella koo SAN fckn	hella flirt hut iono OAKLAND
 New York	NETS KNICKS	BRONX	iam cab	oww	wasssup nm
 Los Angeles	#KOBE #LAKERS AUSTIN	#LAKERS load HOLLYWOOD imm MICKEY TUPAC	omw tacos hr HOLLYWOOD	af *papi* raining th bomb coo HOLLYWOOD	wyd coo af *nada* tacos messin fasho bomb
 Lake Erie	CAVS CLEVELAND OHIO BUCKS od COLUMBUS	premiere prod joint TORONTO onto designer CANADA village burr	stink CHIPOTLE tipsy	;d blvd BIEBER hve OHIO	foul WIZ salty excuses lames officer lastnight

Abbildung 16: Beispiele aus dem geografischen Themenmodell nach Eisenstein *et al.* (61).

Hale *et al.* (26) analysierten die Abhängigkeit zwischen dem Standort und der Sprache von Tweets. Sie untersuchten, ob die im Tweet verwendete Sprache mit der Landessprache der Position des Tweets übereinstimmt. Dabei wurden drei Programme zur Spracherkennung verwendet, Compact Language Detection (CLD), Xerox und Alchemy. Für diese Studie kam die Fleiss‘ Kappa als Metrik zum Einsatz. Die Skala reicht von -1 für eindeutige Ablehnung bis zur 1 für totale Übereinstimmung. Als Datenbasis liegen Tweets aus den Einzugsgebieten der Städte Kairo, Montreal, San Diego und Tokyo vor. Die Zuordnung erfolgt nur anhand des Texts. Ein Vergleich der Sprachklassifikation erfolgte zwischen mehreren Menschen und den drei Programmen (Tabelle 17). Die einzelnen Ergebnisse werden mit der Fleiss‘ Kappa zusammengefasst. Die Übereinstimmung mit dem richtigen Ort ist bei der menschlichen Klassifikation deutlich höher.

Ort	Fleiss‘ Kappa manuell	Fleiss‘ Kappa maschinell
Kairo	**0.839**	0.210
Montreal	**0.759**	0.463
San Diego	**0.867**	0.388
Tokio	**0.854**	0.219
Gesamt	**0.896**	0.481

Tabelle 17: Ergebnisse der Sprachklassifikation nach Hale *et al.* (26)

Verglichen wurden die Ergebnisse von drei Spracherkennungsprogrammen und der Spracheinstellung von Twitter. Als Vergleichswert, ob die Zuordnung zum richtigen Ort durch die Software erfolgte, kamen die Ergebnisse der menschlichen Klassifikation zur Anwendung. Die Tabelle 18 gibt an, zu wie viel Prozent die Erkennung durch die Software mit der menschlichen Spracherkennung übereinstimmt. Die Software Alchemy schnitt drei Mal sehr gut ab, schlug aber bei Tokio fehl. Die Ursache für die fehlerhafte Klassifizierung in Bezug auf Japanisch ist von den Autoren nicht erläutert worden.
Die durchschnittlich beste Klassifikation wurde von CLD vorgenommen.

Ort	Alchemy	Xerox	CLD	Twitter UI Lang
Kairo	**75,4 %**	56,9 %	72,1 %	42,3 %
Montreal	**88,7 %**	75,8 %	81,0 %	73,7 %
San Diego	**92,0 %**	75,1 %	85,3 %	89,7 %
Tokio	56,5 %	63,7 %	**90,4 %**	82,7 %
Gesamt	78,2 %	67,9 %	**82,2 %**	72,1 %

Tabelle 18: Ergebnisse der Sprachklassifikation mit verschiedenen Sprachanalyseprogrammen nach Hale *et al.* (26).

In Bezug auf Blogs untersuchten 2007 Hurst *et al.* (64) die geografische Verteilung von Sprache und Domain von Bloggern. Hierfür entwickelten sie ein System zur Adressextraktion, bei dem die Adressdaten aus der Webseite bzw. aus dem Webblog über ein Muster gefunden werden. Die Adresse des Bloggers wird anschließend mit der Sprachherkunft und der Domainadresse des Blogs verglichen.
Die Untersuchungen ergaben, dass eine statistisch signifikante Verzerrung zwischen dem Herkunftsort des Bloggers und der verwendeten Sprache sowie Domain vorliegt. Es schreiben beispielsweise mehr Blogger in englischer Sprache, als tatsächlich aus englischsprachigen Ländern stammen.

Vergleich der Ansätze:

Die Vorteile bei der Lokalisierung von Tweets anhand des Textes bestehen darin, dass keine weiteren Informationen, wie Standort, Nutzerbeziehungen oder Metadaten, benötigt werden.
Bei Sprachmodellen entfällt auch die Toponymextraktion.

Einschätzung der einzelnen Ansätze:

Autor	Methode	Ergebnisse und Einschätzung	
Hecht *et al.* (58)	Sprachmodell, welches ein MN[20]-Bayes Modell mit einem Wortvektor verwendet.	Mit dem Ansatz ist es möglich, Tweets zu 30,28 % auf das richtige Bundesland zu verorten. Die Lokalisierung auf Bundeslandebene ist zu ungenau.	
Eisenstein *et al.* (61)	Themenmodell, verwendet die Latent Dirichlet Allocation von Blei *et al.* (62).	Tweets können zu 24 % einem Bundesland zugeordnet werden und die Durchschnittsabweichung beträgt 900 km. Die Lokalisierung auf Bundeslandebene ist zu ungenau.	
Hale *et al.* (26)	Spracherkennung und Lokalisierung über Sprache	Es können maximal Tweets zwischen verschiedensprachigen Ländern unterschieden werden. Zur Lokalisierung nicht geeignet.	
Ikawa *et al.* (60)	Sprachmodell, welches regionale Wörter von Ortsmitteilungs-diensten verwendet.	Das Verfahren funktioniert nur in Bezug auf sehr wenige Tweets und ist nur marginal besser als trivialere Ansätze.	
Paradesi (59)	Toponymextraktion aus dem Tweettext.	15,81 % der Tweets werden exakt auf den Ort bestimmt (Precision). Es bedarf eines Tweeters mit validem Ort im Standortfeld und im Tweettext muss ein Ort mit 2 Wörtern enthalten sein. Das Verfahren verortet nur wenige Tweets, diese jedoch exakt.	
Cheng *et al.* (27)	Sprachmodell, welches regionale Wörter verwendet.	51 % der Tweets werden in einem Radius von 161 km um ihre Stadt verortet (Median). 30 % liegen innerhalb eines Radius von 16 Kilometern und die Durchschnittsabweichung beträgt 862 km.	
Kinsella *et al.* (25)	Sprachmodell, welches einer Koordinate einen Korpus an Wörtern zuordnet.	Genauigkeitslevel	Tweets korrekt verortet
		Staat:	53,2 %
		Bundesstaat	31,6 %
		Stadt	29,8 %
		Postleitzahl	13,9 %

Tabelle 19: Übersicht über die Verfahren der Tweetlokalisierung auf Basis des Texts

[20] Multinominales Naives Bayes Modell - Siehe: Data Mining (94).

3.5. Analysen des Standortfelds

Analysen des Standortfelds von Nutzern der Sozialen Medien ergaben, dass regionale Differenzen bezüglich der relativen Häufigkeit von Adresseintragungen existieren. Nach Hurst *et al.* (64) geben beispielsweise Russen häufiger ihre Adresse preis als Chinesen. In einigen amerikanischen Staaten teilen Facebooknutzer ihre Adresse, proportional zur Population, häufiger mit als in anderen Regionen, so das Forschungsergebnis von Backstrom *et al.* (65). Bei den Gewohnheiten der Adressfreigabe auf Facebook existieren keine Altersunterschiede, jedoch geben männliche Nutzer ihre Adresse signifikant häufiger an.Backstrom *et al.* (65) nehmen an, dass es für den Nutzer einfacher ist, das Standortfeld im Profil leer zu lassen, als falsche Informationen einzutragen.

Die nachfolgenden Analysen des Standortfelds von Hecht *et al.* (58) verdeutlichen, dass die Nutzer jedoch falsche Informationen eintragen. Dabei wurde dieses Standortfeld auf die Qualität der enthaltenen Daten und auf dessen geografische Größenordnung untersucht. Die Analyse ergab, dass zu 84 % das Feld ausgefüllt ist, jedoch nur 66 % aller Felder einen geografischen Ort bezeichnen, siehe Abbildung 17. 18 % enthalten nicht verwertbare Angaben, wie *„Justin Biebers heart"* oder *„NON YA BISNESS!!"* .[21]

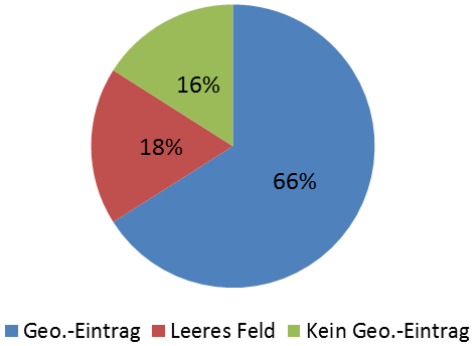


Abbildung 17: Analyse der Standortangaben von Hecht *et al.* (58)

Die verwertbaren Daten lassen sich in die geografischen Klassen der Länder, Regionen, Bundesstaaten, Bezirke, Städte, Stadtbezirke und der genauen Adresse einteilen. In der Abbildung 18 ist zu erkennen, dass die meisten Nutzer mit 64 % eine Stadt eingetragen haben. Nur in 4 % der Fälle enthält der eingetragene Text mehrere Orte. Twitterapps für Smartphones tragen oftmals die GPS Koordinaten automatisch in das Standortfeld ein, dies ist zu 11,5 % der Fall.

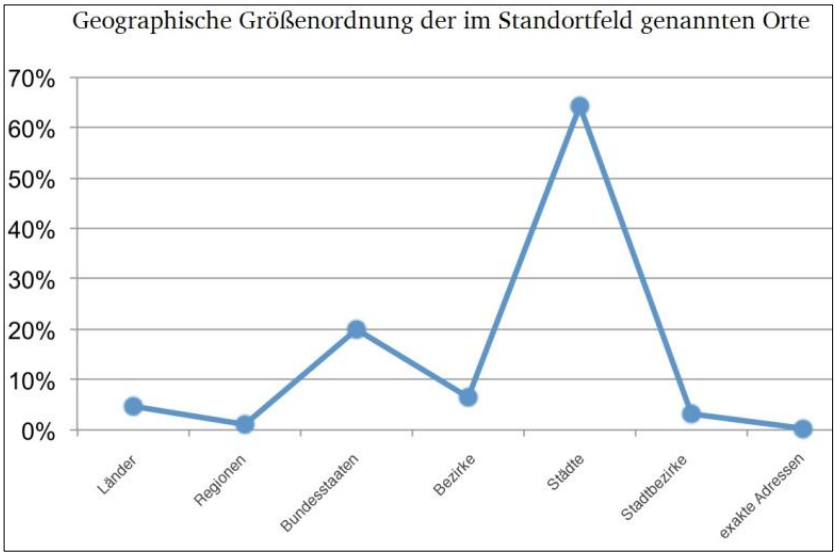


Abbildung 18: Vergleich Hecht *et al.* (58)

[21] Bedeutet ungefähr: *„Geht dich nichts an!"*.

3.6. Lokalisierung auf Basis des Standortfelds

Hale *et al.* (26) analysierten die Beziehung zwischen der Position des Tweets und der des Orts aus dem Standortfeld. Hierfür wurden jeweils 1000 zufällige Tweets aus dem Einzugsgebiet der Städte Kairo, Montreal, San Diego, Tokyo ausgewählt. Dabei kam ein Polygon mit Koordinaten des Stadtgebiets zum Einsatz. Nur Tweets innerhalb des Polygons fanden Verwendung. Alle Tweets verfügen über ein Koordinatenpaar, welches über GeoIP[22] oder GPS-Position gesetzt wurde. Mit den geografischen Lexika von Yahoo und Google erfolgt eine Umwandlung der Zeichenkette des Standortfelds in Koordinaten. Anschließend wird geprüft, ob das Koordinatenpaar innerhalb des Polygons liegt.
In den Tabellen 21 und 22 sind die Einträge dann als fehlerhaft dargestellt, wenn kein oder ein unkorrekter Eintrag ermittelt wurde.

Insgesamt betrachtet, sind ca. 45 % aller Tweets im Polygon der Stadt enthalten.
Ein Vergleich zwischen den beiden Verortungswerkzeugen verdeutlicht, dass Yahoo (ø 4,3 % Fehler) deutlich besser verortet als Google (ø 11,7 % Fehler).

		Yahoo			**Google**		
Stadt	Standortfeld leer	In der Box	Außerhalb der Box	Fehlerhaft	In der Box	Außerhalb der Box	Fehlerhaft
Kairo	20,00 %	43,20 %	33,60 %	**3,20 %**	**44,00 %**	29,80 %	6,20 %
Montreal	13,60 %	46,70 %	35,80 %	**3,80 %**	**56,00 %**	22,80 %	7,60 %
San Diego	14,00 %	**44,40 %**	35,60 %	**6,00 %**	40,40 %	34,70 %	10,90 %
Tokyo	15,80 %	**45,90 %**	34,00 %	**4,30 %**	42,00 %	20,00 %	22,20 %
Gesamt	15,90 %	45,10 %	34,80 %	**4,30 %**	**45,60 %**	26,80 %	11,70 %

Tabelle 20: Übersicht über die Lokalisierung von Tweets auf Basis des Standortfelds nach Hale *et al.* (26)

Hale *et al.* (26) haben außerdem die Distanz zwischen der Position des Tweets und der des Standorts erforscht. Für den Eintrag im Standortfeld erfolgte mittels der geografischen Lexika eine Umwandlung in ein Koordinatenpaar, welches das Zentrum der Stadt repräsentiert.
Anzumerken ist, dass Nutzer, welche mehrmals getwittert haben, auch mehrmals enthalten sind. In der Tabelle 21 sind neben den Ergebnissen der Entfernungsmessung auch die Ergebnisse der Evaluierung der geografischen Lexika dargestellt. Dabei untersuchten die Autoren, wie oft ein richtiger Eintrag gefunden wurde. Auffällig ist, dass die Mittelwerte der Distanz zwischen dem Tweet und dem Standort bezüglich der Städte sehr unterschiedlich sind. Mit dem Yahoo Placefinder hat Kairo einen Mittelwert 494 Meilen und Tokyo von 2780 Meilen. Das deutet darauf hin, dass in der Tweetmenge von Tokyo mehr oder größere Ausreißer enthalten sind. Der Mittelwert ist bei der Verortung mit Google stets geringer als mit dem Yahoo PlaceFinder. In Bezug auf den Median hat Yahoo in nur einem der vier Fälle Google unterboten. Weiterhin ist die Rate der fehlerhaften Verortung mit Google höher, jedoch sind Median und Mittelwert geringer. Schlussfolgernd lässt sich konstatieren, dass Yahoo mehr Einträge richtig zuordnet, jedoch die Ergebnisse mit Google genauer sind.

		Yahoo PlaceFinder				**Google Geocoding API**			
Stadt	Standortfeld leer	Richtiger Eintrag gefunden	Fehlerhaft	Mittelwert in Meilen	Median in Meilen	Richtiger Eintrag gefunden	Fehlerhaft	Mittelwert in Meilen	Median in Meilen
Kairo	20,1 %	**73,7 %**	**6,2 %**	494,98	7,13	70,3 %	9,6 %	**451,98**	**6,975**
Montreal	10,8 %	**86,0 %**	**3,2 %**	933,12	**5,47**	80,0 %	9,2 %	**821,10**	7,876
San Diego	10,5 %	**84,4 %**	**5,1 %**	1.461,89	12,67	76,2 %	13,3 %	**726,80**	**11,359**
Tokyo	16,2 %	**79,2 %**	**4,6 %**	2.780,19	16,30	61,4 %	22,4 %	**502,36**	**6,929**

Tabelle 21: Übersicht über die Ergebnisse der Tweetlokalisierung mit zwei geografischen Lexika nach Hale *et al.* (26)

[22] GeoIP ist die geografische Herkunft einer IP-Adresse.

3.7. Lokalisierung auf Basis der Nutzerbeziehungen

Studien von Kwak *et al.* (66) und Davis *et al.* (63) haben ergeben, dass die Verbindungen der Nutzer untereinander bei Twitter und Facebook unterschiedlich sind. Die Beziehungen zwischen Tweetern sind nach Kwak *et al.* (66) zu 77,9 % einseitig. Eine einseitige Beziehung besteht dann, wenn ein Tweeter einem anderem folgt (followed), dieser jedoch die Beziehung nicht erwidert. Twitter wird vielmehr als eine Informationsquelle verwendet. Der Charakter von Facebook liegt mehr in der Natur der Sozialen Netzwerke.
McGee *et al.* (67) erforschten die Korrelation zwischen den sozialen Beziehungen und der geografischen Distanz von Twitternutzern. Tweeter, zwischen denen eine starke gegenseitige Verbindung existiert, sind geografisch näher zueinander zu verorten. Abbildung 19 gibt die Distanz zwischen den Nutzern für verschiedene Beziehungsgrade an. In der Tabelle 22 sind die Beziehungsgrade nach McGee *et al.* (67) erläutert.

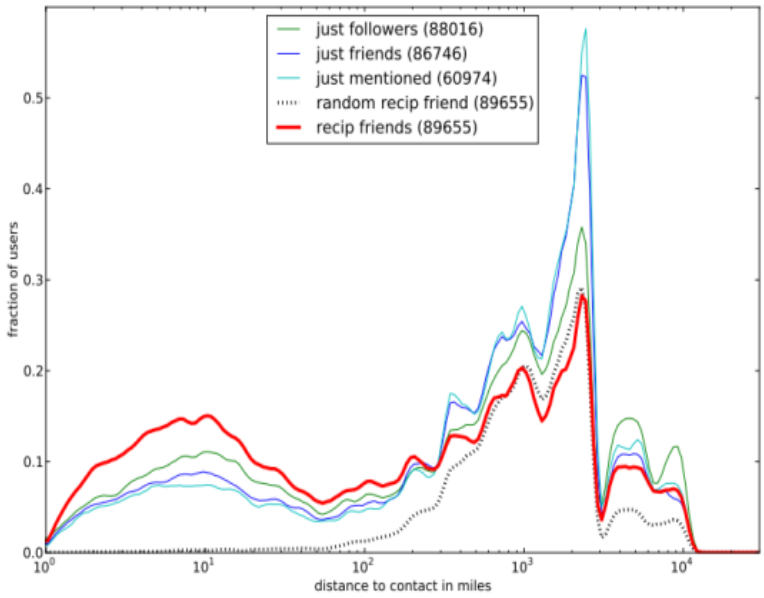


Abbildung 19: Ergebnisse von McGee *et al.* (67)

Beziehung	Erläuterung
just follower	The geo-located user is followed by this user, but does not follow them.
just friend	The geo-located user follows this user and is not followed back.
just mentioned	The users do not follow each other, but the geo-located user mentioned the name of the other user in a tweet.
reciprocal friend	The geo-located user follows this user and is followed back.
random recip friends	The distance between random, unrelated users.

Tabelle 22: Erläuterungen zu den Beziehungen in Abbildung 19 nach McGee *et al.* (67)

Nach Scellato (68) sind 40-50 % aller befreundeten Nutzer von Ortsmitteilungsdiensten (Foursquare, Brightkite, Gowalla) im Umkreis von 100 km zueinander angesiedelt, in 3 % der Fälle sogar im Radius von einem Kilometer. Backstrom *et al.* (65) entwickelten eine Funktion, welche die Entfernung und die Beziehungswahrscheinlichkeit von Twitternutzern miteinander verbindet. Hierfür wurden die unterschiedlichen geografischen Bevölkerungsdichten analysiert. Abbildung 20 führt den Nachweis, dass die Wahrscheinlichkeit, miteinander befreundet zu sein, höher ist, je geografisch näher die Nutzer zueinander angesiedelt sind. Weiterhin ist in Gebieten mit einer geringen Bebauungsdichte die Wahrscheinlichkeit höher, dass die Nutzer geografisch näher zueinander verortet sind als in stark bebauten Gebieten.

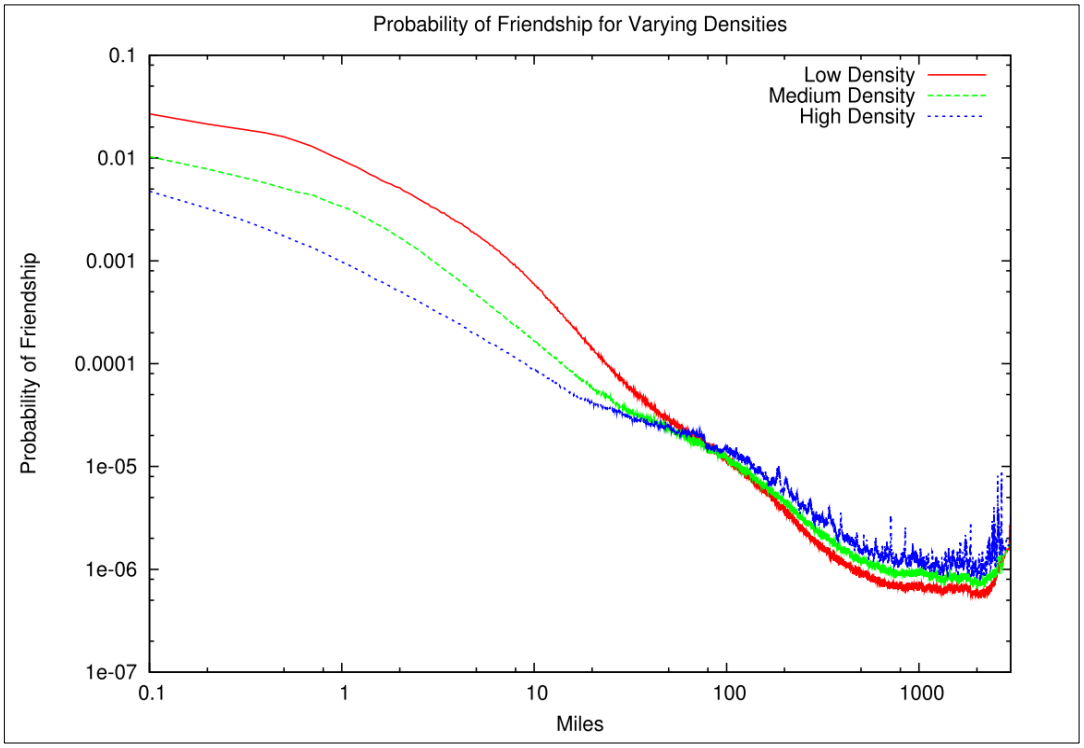


Abbildung 20: Diagramm Freundschaftswahrscheinlichkeit in Bezug zur Entfernung zwischen den Freunden Backstrom *et al.* (65)

Davis *et al.* (63) wenden ein Verfahren an, um den Standort des Twitternutzers anhand seiner Beziehungen zu bestimmen. Analysen haben nachgewiesen, dass Nutzer mit zu wenigen oder zu vielen Kontakten ungeeignet sind. Bei zu wenigen Kontakten sind diese Nutzer oftmals inaktiv und bei Tweetern mit einer zu großen Anzahl von Followern handelt es sich in der Regel um bekannte Persönlichkeiten oder Institute.
Für die Bestimmung des Standorts werden aus den Profilen der befreundeten Twitternutzer deren Position ausgelesen. Eine einfache Möglichkeit, den Ort zu bestimmen, ist die Wahl des am häufigsten vorkommenden Orts unter den Nutzerbeziehungen. Damit sind 45 % der prognostizierten Standorte identisch mit dem aus dem Standortfeld.

Backstrom *et al.* (65) bestimmen den Standort der Facebooknutzer anhand eines Graphen, welcher aus den Standorten der befreundeten Nutzer besteht. Der triviale Ansatz ist, den Schwerpunkt aus den Orten zu wählen, dabei wird ein Polygon mit den Orten der Beziehungen aufgespannt und der Punkt mit der kürzesten Entfernung zu allen Orten ermittelt. Die Analyse dieses Ansatzes ergab, dass dieser aufgrund der Streuung der Beziehungen nicht geeignet ist. Beispielsweise hat ein Darmstädter Nutzer 15 Freunde in der Umgebung von Darmstadt und einen Freund in Blacksburg (USA). Die Durchschnittsentfernung liegt weit entfernt vom eigentlichen Standort des Nutzers.

Der von Backstrom *et al.* (65) entwickelte Graphenalgorithmus berechnet die Standortkoordinate aus den verschiedenen Kantenwahrscheinlichkeiten. Zu jedem befreundeten Kontakt mit einem bekannten Standort wird mit einer Entfernungsrankingfunktion die Kantenwahrscheinlichkeit ermittelt. Die Rankingfunktion ist eine Regressionsfunktion, welche unterschiedliche Populationsdichten und deren Wahrscheinlichkeit, über eine Distanz befreundet zu sein, aggregiert. Innerhalb eines Radius von 40 km können Nutzer mit mehr als 16 georteten Freunden zu 69,1 % lokalisiert werden. Zum Vergleich sind die Ergebnisse der Lokalisierung über die IP-Adresse des Nutzers anzuführen, welche 57,2 % innerhalb des Radius verortet.

Vergleich der Ansätze:

Mit den Ansätzen auf Basis der Beziehungen eines Tweeters lässt sich dessen Heimatstandort ermitteln, nicht jedoch die exakte GPS-Position des einzelnen Tweets.

Autor	**Methode**	**Ergebnisse und Einschätzung**
Davis *et al.* (63)	Häufigster Ort unter den Beziehungen	45 % der prognostizierten Orte sind identisch mit dem Ort aus dem Standortfeld. Nur zur Lokalisierung des Standorts des Tweeters geeignet.
Backstrom *et al.* (65)	Berechnung des Schwerpunkts aus den Orten der Beziehungen.	Ungeeignet zur Lokalisierung des Standorts und der von Tweets. Ein Beispiel verdeutlicht die Ungeeignetheit. Hat ein Tweeter drei Kontakte in Darmstadt und einen in den USA, dann liegt der ermittelte Schwerpunkt wahrscheinlich über dem Atlantik.
Backstrom *et al.* (65)	Graphen-algorithmus	Innerhalb eines Radius von 40 km können Nutzer mit bis zu 69,1 % lokalisiert werden. Nur zur Lokalisierung des Standorts geeignet.

Tabelle 23: Übersicht über die Verfahren der Standortlokalisierung des Nutzers

3.8. Metainformationen

Aus dem Nutzerprofil lassen sich über die Twitter APIs weitere Metainformationen gewinnen. Diese Informationen werden automatisch beim Besuch der Twitterplattform über den Browser oder über die Mobil-Geräte übermittelt. Somit liegen Informationen über die Zeitzone[23] und das UTC[24]-Offset des Tweeters vor und können zu seiner Verortung genutzt werden. Krishnamurthy *et al.* (69) analysierten 2008 die Verteilung der Twitternutzer über die Bereiche der Weltuhrzeit. Mithilfe der Grafiken in der Abbildung 21 ist zu erkennen, dass die häufigsten Nutzer im Offsetbereich der Ostküste und Westküste der USA, in Europa sowie in Westasien und in Australien beheimatet sind[25].

Hale *et al.* (26) untersuchten die Korrelation zwischen der Zeitzone und der GPS-Position des Tweets. Analog dazu erfolgte die Untersuchung zum UTC-Offset. Durchschnittlich ist die Zeitzone zu 64,4 % identisch mit der des Tweets. Am Beispiel von Montreal mit 41,7 % und Tokyo mit 57 % wird deutlich, dass Abweichungen feststellbar sind.
Bei dem Vergleich der Ergebnisse muss darauf geachtet werden, dass die Fläche des Offsets um ein Vielfaches größer ist als die der Zeitzone. Somit fallen mehr Ausreißer in das UTC Offset.

Stadt	Anzahl	Korrekte Zeitzone	Korrektes UTC-Offset
Kairo	1.952	**54,6 %**	54,4 %
Montreal	5.235	41,7 %	**57,0 %**
San Diego	9.292	57,1 %	**60,5 %**
Tokyo	55.573	68,2 %	**72,3 %**
Gesamt	72.052	64,4 %	**69,2 %**

Tabelle 24: Ergebnisse der Untersuchung von Zeitzone und UTC-Bereich und der Position des Tweets nach Hale *et al.* (26)

Die Zeitzone und das UTC-Offset können aus zwei Gründen von der des Tweets abweichen:

- der Tweeter befindet sich nicht am Standort
- die Eintragung von Zeitzone/UTC-Offset ist nicht korrekt.

Allein mit diesem Ansatz sind Tweets nur sehr ungenau lokalisierbar. Die Zeitzone kann als Mittel zur Disambiguierung verwendet werden.

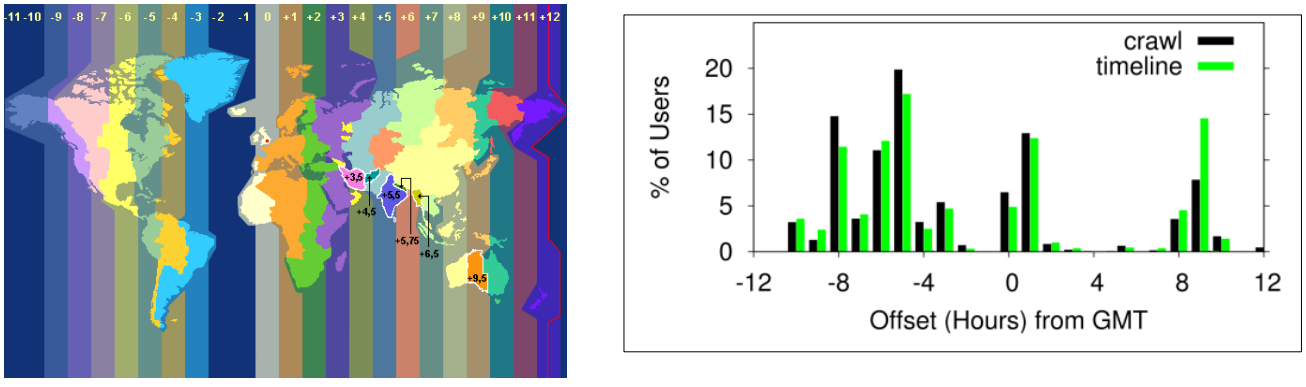


Abbildung 21: Rechts: Weltkarte mit eingezeichneter Weltuhrzeit (UTC)
Links: Verteilung der Tweets über die Weltuhrzeitbereich nach Krishnamurthy *et al.* (69)

[23] Zeitzonen: Bereich in der dieselbe Uhrzeit gilt. Vgl. Brockhaus Band 30, S. 516 .

[24] UTC: Universal Time Coordinated – koordinierte Weltuhrzeit. Vgl. Brockhaus (93) Band 30, S. 516 und Band 28, S. 483.

[25] „Crawl" und „Timeline" bezeichnen zwei Datensets, auf deren Grundlage die Untersuchung durchgeführt wurde.

3.9. Weitere Ansätze

Twitternachrichten beinhalten zu 22 % eine URL (Boyd *et al.* (30)), welche auf Webseiten oder Fotos verweisen. Ein bekannter Dienst zum Speichern sowie Teilen von Bildern und Fotos ist *Flicker* (70). Bei diesem besteht die Möglichkeit, Fotos mit einer Geoposition auszustatten. Hecht *et al.* (57) untersuchten die Entfernung zwischen lokalisierten Fotos und der Eintragung im Standortfeld des Flickernutzers. Innerhalb eines Radius von 100 km waren 53 % der Fotos und Bilder angesiedelt.

McCurley (71) widmete sich der Extraktion und der Lokalisierung von Texten anhand der darin enthaltenen Informationen, wie Telefonnummern, Adressen und Postleitzahlen. Enthaltene Hyperlinks, welche einen geografischen Inhalt besitzen, können zur Verortung herangezogen werden.

3.10. Vergleich der Ansätze

Dieser Abschnitt beinhaltet einen Vergleich der vorgestellten Ansätze. Es werden hier nur diejenigen Verfahren einander gegenübergestellt, welche in der Twitterdomäne anwendbar sind.
Die Ergebnisse der einzelnen Ansätze lassen sich nicht direkt miteinander vergleichen, da die Datenbasen, auf deren Grundlage die Verfahren getestet wurden, unterschiedliche sind. Das betrifft die Menge, den Erhebungszeitraum und die Erhebungsmethode der Daten. Unter der Erhebungsmethode ist der Umfang des Datenstreams zu verstehen, welcher unterschiedlich große Anteile am Gesamtvolumen wiedergibt. Die Rohdaten werden anschließend gefiltert. Paradesi (59) arbeitet ohne Filterung, betrachtet nur die Phrasen, welche mindestens zwei Wörter beinhalten. Andere Autoren filtern die User mit zu geringer und zu hoher Tweetanzahl. Es kommen unterschiedliche geografische Lexika zum Einsatz, welche sich im Hinblick auf Qualität und Umfang differenzieren. Weiterhin existiert keine einheitlich genormte Metrik zur Evaluierung der Systeme. Somit ist kein objektiver Vergleich der Verfahren möglich.

Zusammenfassung der Methoden zur Lokalisierung von Tweets:

Ansatz/Autor	**Methode**	**Jeweils beste Ergebnisse**		**Benötigte Info. zur Lokalisierung**
Auf Textbasis/ Paradesi (59)	Toponymextraktion aus dem Tweettext.	15,81 % der Tweets werden exakt auf den Ort bestimmt (Precision).		Tweeter mit validem Ort im Standortfeld. Im Text des Tweets muss ein Ort mit 2 Wörtern enthalten sein.
Auf Textbasis/ Cheng *et al.* (27)	Sprachmodell, welches regionale Wörter verwendet.	51 % der Tweets werden im Radius von 161 km um ihre Position verortet (Median). Durchschnittsabweichung: 862 km		Keine Einschränkung
Auf Textbasis/ Kinsella *et al.* (25)	Sprachmodell, welches einer Koordinate einen Korpus an Wörtern zuordnet.	Genauigkeitslevel	Tweets korrekt verortet	Keine Einschränkung
		Staat: Bundesstaat Stadt Postleitzahl	53,2 % 31,6 % 29,8 % 13,9 %	

Tabelle 25: Zusammenfassung der Methoden zur Lokalisierung von Tweets – Erster Teil

<table>
<tr><th>Ansatz/ Autor</th><th>Methode</th><th>Jeweils beste Ergebnisse</th><th>Benötigte Info. zur Lokalisierung</th></tr>
<tr><td>Auf Textbasis/ Hecht et al. (58)</td><td>Sprachmodell, welches ein MN[26]-Bayes Modell mit einem Wortvektor verwendet.</td><td>Lokalisierung der Tweets auf maximal Bundeslandebene.<table><tr><td>Geografische Größenordnung</td><td>Genauigkeit (Referenz/Baseline)[27]</td></tr><tr><td>Land-Uniform</td><td>72,71% (25%)</td></tr><tr><td>Land-Random</td><td>88,86% (82,08%)</td></tr><tr><td>Bundesland-Uniform</td><td>30,28% (5,56%)</td></tr><tr><td>Bundesland-Random</td><td>27,31% (15,06%)</td></tr></table></td><td>Keine Einschränkung</td></tr>
<tr><td>Auf Textbasis/ Eisenstein et al. (61)</td><td>Themenmodell, verwendet die Latent Dirichlet Allocation von Blei et al. (62).</td><td>24 % aller Tweets können einem Bundesstaat zugeordnet werden. Durchschnittsabweichung: 900km</td><td>Keine Einschränkung</td></tr>
<tr><td>Auf Textbasis/ Hale et al. (26)</td><td>Spracherkennung und Lokalisierung über Sprache</td><td>Es kann maximal zwischen verschiedensprachigen Ländern differenziert werden.</td><td>Keine Einschränkung</td></tr>
<tr><td>Auf Textbasis/ Ikawa et al. (60)</td><td>Sprachmodell, welches regionale Wörter von Ortmitteilungs-diensten verwendet.</td><td>Der Ansatz hat einen sehr geringem Recall und ist nur marginal besser als der triviale Ansatz, immer die häufigste vergangene Position des Tweets zu verwenden (Baseline).</td><td>Nur Orte, welche in Ortsmitteilungs-diensten enthalten sind.</td></tr>
<tr><td>Auf Basis des Standortfelds/ Hale et al. (26)</td><td>Vergleich der Position des Tweets und Standortfeld-eintragung</td><td>45,6 % der Tweets liegen im Polygon der im Standortfeld angegebenen Stadt. Weitere Ergebnisse:<table><tr><td>Stadt</td><td>Mittelw.</td><td>Median</td></tr><tr><td>Kairo</td><td>452,0 mi</td><td>6,975 mi</td></tr><tr><td>Montreal</td><td>821,1 mi</td><td>7,876 mi</td></tr><tr><td>San Diego</td><td>726,8 mi</td><td>11,359 mi</td></tr><tr><td>Tokyo</td><td>502,4 mi</td><td>6,929 mi</td></tr></table></td><td>Tweeter mit validem Ort im Standortfeld</td></tr>
<tr><td>Auf Metadaten-basis Hale et al. (26)</td><td>Zeitzone/ UTC-Offset</td><td>Verortung auf Zeitzonen- oder Länderebene. Durchschnittlich stimmen die Zeitzone mit 64,4 % und das UTC-Offset mit 69,2 % mit der des Tweets überein.</td><td>Keine Einschränkung</td></tr>
</table>

Tabelle 26: Zusammenfassung der Methoden zur Lokalisierung von Tweets – Zweiter Teil

[26] Multinominaler Naives Bayes Modell.

[27] Baseline: Wahrscheinlichkeit bei zufälliger Wahl der Tweets in die vorgegebenen Klassen.

Zusammenfassung zur Lokalisierung des Standortes[28] :

Ansatz/ Autor	Methode	Jeweils beste Ergebnisse	Benötigte Info. zur Lokalisierung
Auf Basis der Beziehungen/ Davis *et al.* (63)	Häufigster Ort unter den Beziehungen	Bei 45 % der Nutzer entspricht der prognostizierte Ort dem aus dem Standortfeld.	Tweeter benötigt ausreichende Beziehungen zu anderen Tweetern.
Auf Basis der Beziehungen/ Backstrom *et al.* (65)	Wahl des Orts mittels eines Graphen-algorithmus	69,1 % können im Radius von 40 km um ihre tatsächliche Position verortet werden.	Tweeter benötigt ausreichende Beziehungen zu anderen Tweetern.

Tabelle 27: Zusammenfassung der Methoden zur Lokalisierung des Standortes

[28] Siehe Definition Kapitel 1.2.

4. Umsetzung der Tweetlokalisierung auf Basis des Standortfelds

Die folgenden Abschnitte beschreiben die Herausforderungen der Lokalisierung auf Basis des Standortfelds, den gewählten Ansatz und die Implementierungsdetails des Prototyps zur Evaluierung des Modells.

4.1. Herausforderungen der Ortsextraktion und Disambiguierung aus dem Standortfeld

Evaluation des Standortfelds

Für die Analyse des Standortfelds wurden 1000 zufällige Twitternutzer mit der Spracheinstellung *deutsch* im Profil ausgewählt. Bei 23,3 % der Nutzer ist dieses Feld nicht gesetzt. Die Standortangaben lassen sich in sieben Klassen einteilen, siehe Tabelle 28. Bei der Klassifikation wurden Ortsangaben, welche Abkürzungen enthielten, doppelt zugeordnet.

Klasse	**Klassenbeschreibung**	**Beispiel aus dem Standortfeld**	**Relative Häufigkeit**
Stadt	Enthält nur Städte.	*„Köln“*	35,2 %
Land	Beinhaltet nur Länder oder Bundesländer.	*„Deutschland“*	22,1 %
Eindeutig	Identifiziert einen Ort eindeutig.	*„Ingelheim, Deutschland“,* *„Frankfurt am Main ♥“,* *„Moltkestr. 51, 12203 Berlin“*	20,2 %
Kein geografischer Bezug	Enthält keine geografischen Daten.	*„Delirium“,“ „Hogwarts ♥“, „earth :]“,* *„Gerɯαɳy♥♥“*[29]	15,1 %
Abkürzung	Enthält Abkürzungen.	*„D-Town“, „drtmnd“*	4,9 %
Mehrfache Orte	Mehrere Ortsangaben im Standortfeld	*„mainz, münchen oder dazwischen“*	2,8 %
Koordinaten	Enthält GPS-Koordinaten	*„iPhone: 48.156509,11.502149“,* *„N 51°27' 0" / E 6°34' 0"“*	2,5 %

Tabelle 28: Analysen des Standortfelds

Texte innerhalb der Twitterdomäne sind gekennzeichnet durch ihre Kürze, die Verwendung von Umgangssprache und mehrdeutigen Abkürzungen. Bei der Analyse des Standortfelds fällt auf, dass nur wenige Rechtschreibfehler vorkommen, jedoch auch, dass die Großschreibung der Ortsnamen häufig nicht erfolgt. 4,9 % der Eintragungen des Standortfelds beinhalten Akronyme, welche oft Ortsangaben in nicht ISO 3166-Norm (47) enthalten. Einträge ohne scheinbaren geografischen Bezug, wie „Middle Earth“, „couch“ und „Justin Bieber´s heart“, kommen zu 15,1 % vor. Diese Eintragungen sind oftmals durch aktuelle gesellschaftliche Trends und Themen dominiert. Herausfordernd ist, dass diese Angaben reale Orte bezeichnen können. So ist „Middle Earth“ eine Stadt in Maryland (USA) und zu „Couch“ existieren mehrere Orte.
Bei der Disambiguierung sind die in Kapitel 3.3. vorgestellten Verfahren von Amitay *et al.* (33) und von Smith *et al.* (33) nicht verwendbar, da die Texte in Tweets und im Standortfeld zu kurz sind und keine wohlgeformte Sprache bieten.

[29] „Gerɯαɳy♥♥“ ist in schwer maschineninterpretierbarer Schriftart geschrieben.

4.2. Modell

In diesem Abschnitt werden das Modell beschrieben, die verwendeten Filtermöglichkeiten vorgestellt und die Strategien zur Ortsauswahl aus der Ergebnisliste bei der Suchanfrage mit Gisgraphy erläutert.

4.2.1. Modellbeschreibung

Ziel des vorgestellten Modells ist es, Tweets ohne GPS-Koordinatenpaar oder Placelabel zu verorten. Im Rahmen dieser Arbeit erfolgt die Lokalisierung von Tweets anhand der im Nutzerprofil eingetragenen Informationen. Dabei wird der Ort aus dem Standortfeld als Approximation der Position des Tweets angenommen. Das Verfahren beruht auf der Annahme, dass im Standortfeld geografische Eintragungen vorhanden sind und dass sich der Tweeter überwiegend in der Nähe des eingetragenen Orts befindet. Die Abbildung 22 (links) beschreibt vereinfacht das Modell. Der Input ist der zu lokalisierende Tweet mit dem Standort aus dem Nutzerprofil des Tweeters. Als Ergebnis weist das Verfahren dem Tweet ein eindeutiges Koordinatenpaar zu.

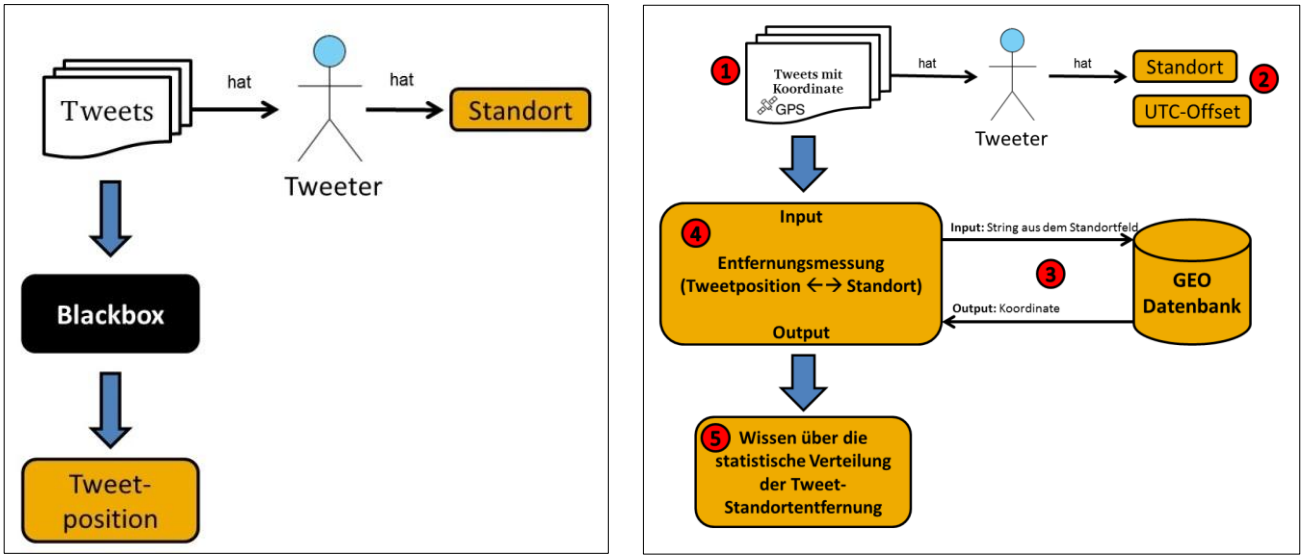


Abbildung 22: Links: Vereinfachte Darstellung des Modells
Rechts: Erweitertes Modell zur Evaluierung der Approximation der Tweetposition

Das Ziel ist es, den Zusammenhang zwischen der Tweetposition und dem Ort aus dem Standortfeld zu ermitteln. Dadurch ist es möglich, die Genauigkeit der Prognose hinsichtlich der Position des Tweets zu untersuchen. Hierfür wird die Abweichung der Prognose von der tatsächlichen Position des Tweets gemessen. Abbildung 24 (rechts) beschreibt den Ablauf. Deshalb verwendet der Algorithmus zur Ermittlung der Distanzverteilung Tweets mit GPS-Koordinaten [1]. Weiterhin finden das Standortfeld und das UTC-Offset aus dem Nutzerprofil [2] des Tweeters Verwendung.
Die Zeichenkette des Standortfelds wird mit einem geografischen Lexikon in Koordinaten umgewandelt, [3] und es erfolgt eine Distanzmessung [4] zum Koordinatenpaar des Tweets. Nach einer statistischen Analyse sind Aussagen über die Entfernungsverteilung möglich [5].

Abbildung 23 erklärt detailliert die einzelnen Schritte des Algorithmus. Zuerst wird aus dem Standortfeld die Zeichenkette ausgelesen [6]. Diese kann Ortsbezeichner, Koordinaten und nicht geografische Inhalte enthalten oder nicht gesetzt sein. Anschließend erfolgt die Extraktion der Koordinaten [7] aus der Zeichenkette, welche direkt den Tweets zugeordnet werden.

Für die Umwandlung der Zeichenketten in Koordinatenpaare kommt ein geografisches Lexikon [8] zur Anwendung. Als Rückgabe erhält man eine Liste von Orten. Diese Ergebnisliste kann jetzt gefiltert werden [9]. Es stehen mehrere Filterungsmöglichkeiten zur Verfügung, die der nächste Abschnitt erläutert.
Bei der Auswahl eines Orts aus der Ergebnisliste [10] und der Auflösung der Mehrdeutigkeiten von Ortsbezeichnern kann das UTC-Offset nützlich sein. Zur Wahl eines Orts aus der Ergebnisliste existieren mehrere Strategien. Deren Vorstellung erfolgt im Abschnitt 4.2.3. Als Ergebnis des vorgestellten Modells kann jedem Tweet mit einer geografischen Angabe im Standortfeld ein Koordinatenpaar [11] zugeordnet werden.

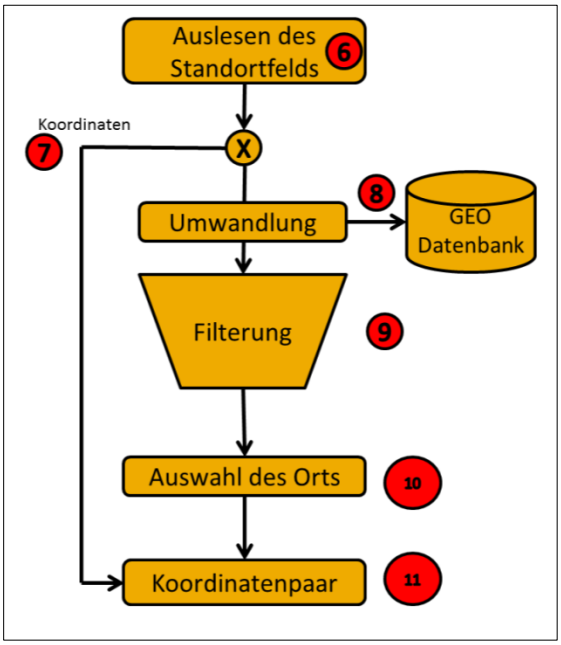


Abbildung 23: Ablaufschritte des Modells

4.2.2. Filterung

Dieser Abschnitt stellt die entwickelten Filterungsmöglichkeiten vor. Die Intension besteht darin, die Genauigkeit der Approximation zu verbessern und eine separierte Betrachtung der unterschiedlichen geografischen Größenordnungseintragungen[30] des Standortfelds zu ermöglichen. Beispielweise können somit Aussagen getroffen werden, wie sich Länderangaben im Standortfeld auf die Präzision der Prognose auswirken.
Bei den Filterungen ist zu erwarten, dass diese sich positiv auf die Genauigkeit auswirken, dass die Anzahl (Recall) der lokalisierenden Tweets jedoch geringer ist.

Entworfene Filterungsmöglichkeiten:

Einschränkung der Suche auf Städte
In Kapitel 4.1. wurde das Standortfeld evaluiert und nachgewiesen, dass eine Ortsangabe in Form einer Stadt am häufigsten eingetragen ist. Um die Suche zu verbessern, erfolgt diese ausschließlich nach Städten.
Hierbei werden Länder automatisch gefiltert. Zu beachten ist, dass Ländernamen, welche gleichzeitig Städtenamen sind, als Stadt interpretiert werden. Granulare Ortsangaben können verloren gehen.

Filterung von Ländern & größeren Arealen
Dieser Filter verwirft geografische Angaben, wie Länder, Kontinente und Ozeane. Analog zum Filter zuvor werden jetzt Städtenamen, welche auch Länder bezeichnen, als Länder interpretiert.

Filterung von Ländern & größeren Arealen und Einschränkung der Suche auf Städte
Um die Herausforderung der Ortsbezeichner, welche simultan Länder und Städte bezeichnen, zu lösen, erfolgt eine Kombination der zuvor vorgestellten Filter.

1. Filterung der Länder & größere Areale
2. Einschränkung der Suche auf Städte

[30] Kapitel 4.1. evaluierte die Standortfeldangaben und im Kapitel 3.5. sind die geografischen Größenordnungen des Standortfelds verzeichnet.

Einschränkung auf Koordinaten
Mit diesem Filter ist es möglich, ausschließlich die Genauigkeit der Prognose für Eintragungen in Koordinatenform zu eruieren. Die gefilterte Menge enthält nur solche Tweets, für die der Tweeter im Standortfeld ein Koordinatenpaar eingetragen hat. Damit entfällt eine Suche im geografischen Lexikon. Die Koordinaten sind vermutlich überwiegend automatisch durch mobile Geräte gesetzt worden.

4.2.3. Auswahl eines Orts aus der Ergebnisliste

Bei der Suche mit Gisgraphy erhält man eine Menge an Orten als Ergebnis. Die Ergebnismenge wird als Liste dargestellt. Diese ist nach einem internen *Scorewert* sortiert, wobei das genaue Sortierverfahren nicht bekannt ist. Es ist erforderlich, aus der Ergebnisliste einen Ort auszuwählen.

Es werden die entwickelten Strategien zur Ortsauswahl vorgestellt:

Random
Wahl eines Orts per Zufall.

Feste Listenposition
Aus der Ergebnisliste wird immer ein Ort in einer bestimmten Position ausgewählt. Dabei ist die Auswahlposition bei allen Suchanfragen fixiert. Beispielsweise wird immer die dritte Position aus der Ergebnisliste gewählt. Da Gisgraphy bereits vorsortiert, ist anzunehmen, dass das wahrscheinlichste Ergebnis an der ersten Position steht. Unter diesem Aspekt ist die Wahl des ersten Listeneintrags zu empfehlen.

Iterative Fuzzy Suche
Unter einer Fuzzy Suche ist eine Ähnlichkeitssuche zu verstehen. Dabei werden ähnliche Schreibweisen gefunden. Erhält ein Suchbegriff keine oder zu wenige Treffer, erfolgt eine Suche nach ähnlicheren Orten.

Majority Voting
Es wird der am häufigsten vorkommende Ort aus der Liste selektiert.
Beispiel: Die Rückgabewerte sind „Darmstadt, Arnstadt, Darmstadt, Kreis Darmstadt, Darmhausen“. Der am häufigsten vorkommende Ort und somit der gewählte ist Darmstadt.

Zeitzone
Nur Orte, welche innerhalb einer Zeitzone liegen, stehen zur Auswahl.

UTC-Offset Bereich
Analog zur Zeitzone stehen nur die Orte zur Auswahl, welche innerhalb des Bereichs derselben Weltuhrzeit liegen.

Population
Der Ort mit der größten Bevölkerung wird gewählt.

Exakte Zeichenkette
Es stehen nur diejenigen Orte zur Auswahl, welche exakt die gleiche Zeichenkette im Standortfeld eingetragen haben.

4.3. Implementierung des Prototyps

In diesem Abschnitt wird die konkrete Implementierung des Prototyps beschrieben. Die Verarbeitung lässt sich in drei elementare Schritte differenzieren, die *Bereitstellung der Datenbasis*, die *Distanzermittlung & Filterung* und die *statistische Auswertung*.

Datenbereitstellung

Das Ziel ist es, für die weitere Verarbeitung und Evaluation eine einheitliche Datenbasis aufzubauen, d.h. für die Experimente immer dieselben Daten zur Verfügung zu stellen. Nach dem Datenbereitstellungsvorgang liegen diese in einer internen Datenstruktur vor, die eine schnelle Verarbeitung gewährleistet. Die Datenstrukturen werden in der Tabelle 29 erklärt.

Name	Erklärung
Tweet	Repräsentiert einen einzelnen Tweet mit den dazugehörigen Informationen, wie der GPS-Koordinate (optional) und dem geschriebenem Text.
TweetSeries	Ist eine Liste von *Tweets*.
Tweeter	Repräsentiert einen einzelnen Tweeter mit Informationen aus dessen Nutzerprofil, wie Name, Standortfeldeintragung und UTC-Offset.
TweeterSeries	Ist eine Datenstruktur zum Aufbewahren von *Tweetern*.

Tabelle 29: Die wichtigsten Datenstrukturen im Überblick

Ablaufschritte:

Aus einer SQL Datenbank werden die Daten von Twitter ausgelesen. Das erhaltene SQL-Resultset wird in die eigenen Datenstrukturen *Tweet* und *Tweeter* verarbeitet, welche in *TweetSeries* und *TweeterSeries* aufbewahrt werden. Anschließend vollzieht sich eine Serialisierung der Serien. Die Abbildung 24 visualisiert die Ablaufschritte.

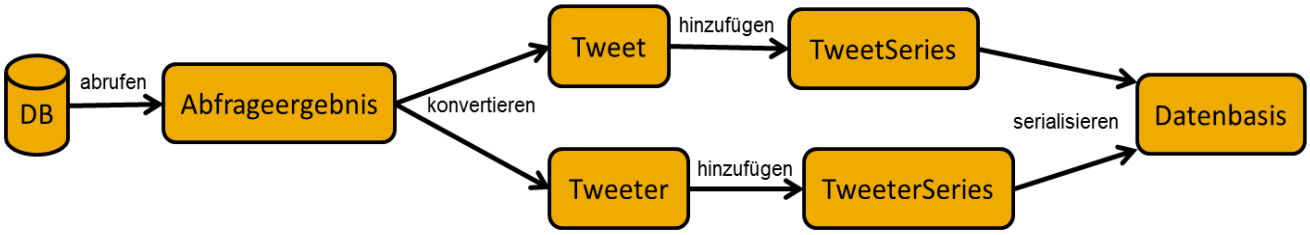


Abbildung 24: Ablaufschritte bei der Erstellung der Datenbasis

Distanzermittlung & Filterung

Für die Ermittlung der Entfernung zwischen Tweet und Standort wird zuerst eine Deserialisierung der Datenbasis vorgenommen. Für alle Tweets wird die Zeichenkette des Standortfelds auf enthaltene Koordinaten[31] hin untersucht und diese mittels Regular Expression extrahiert. Enthält die Zeichenkette Orte, so wird eine Suche in dem geografischen Lexikon Geonames durchgeführt.

[31] Koordinaten vom Typ: Dezimalgrad, Beispiel 48.156509, 11.502149.

Für den Zugriff auf die Geonamesdaten wurde Gisgraphy[32] ausgewählt, weil es sich um eine umfangreiche, kostenfreie und lokal installierbare Opensource Software handelt. Als Rückgabeergebnis der Anfrage mit Gisgraphy erhält der Nutzer eine Liste mit möglichen Orten und dazugehörigen Koordinatenpaaren sowie die Zeitzonen der Orte. Im nächsten Schritt können optional Filter eingesetzt werden. Die Filterung nutzt generell die Möglichkeit, bei der Suche mit Gisgraphy Suchoptionen zu verwenden, speziell die Sucheingrenzung auf einen Ortstyp. Die Funktionsweise der Filterung wird am Beispiel des Länderfilters verdeutlicht. Erhält das System bei der Suche mit Ortstyp *Land* ein Ergebnis zurück, so wird der Suchbegriff als Land klassifiziert.
Der Abgleich, ob der Ort innerhalb des UTC-Offset Bereichs aus dem Nutzerprofil des Tweeters liegt, wird folgendermaßen vorgenommen: Die UTC-Zeitangaben werden in Zeitzonen umgewandelt.
Hierfür wird zu jeder UTC-Zeitangabe eine Liste mit möglichen Zeitzonen erstellt. Anschließend wird ein Abgleich zwischen der Zeitzone des Orts und der Zeitzonenliste durchgeführt. Orte, welche nicht innerhalb der Zeitzonenliste liegen, verwirft das Verfahren.

Es folgt die Distanzermittlung zwischen den Koordinaten des Tweets und denen der gefundenen Orte. Die Ergebnisse werden in einer Liste festgehalten und anschließend zur Speicherung oder Ausgabe weitergeleitet. Die Abbildung 25 visualisiert den beschriebenen Vorgang.

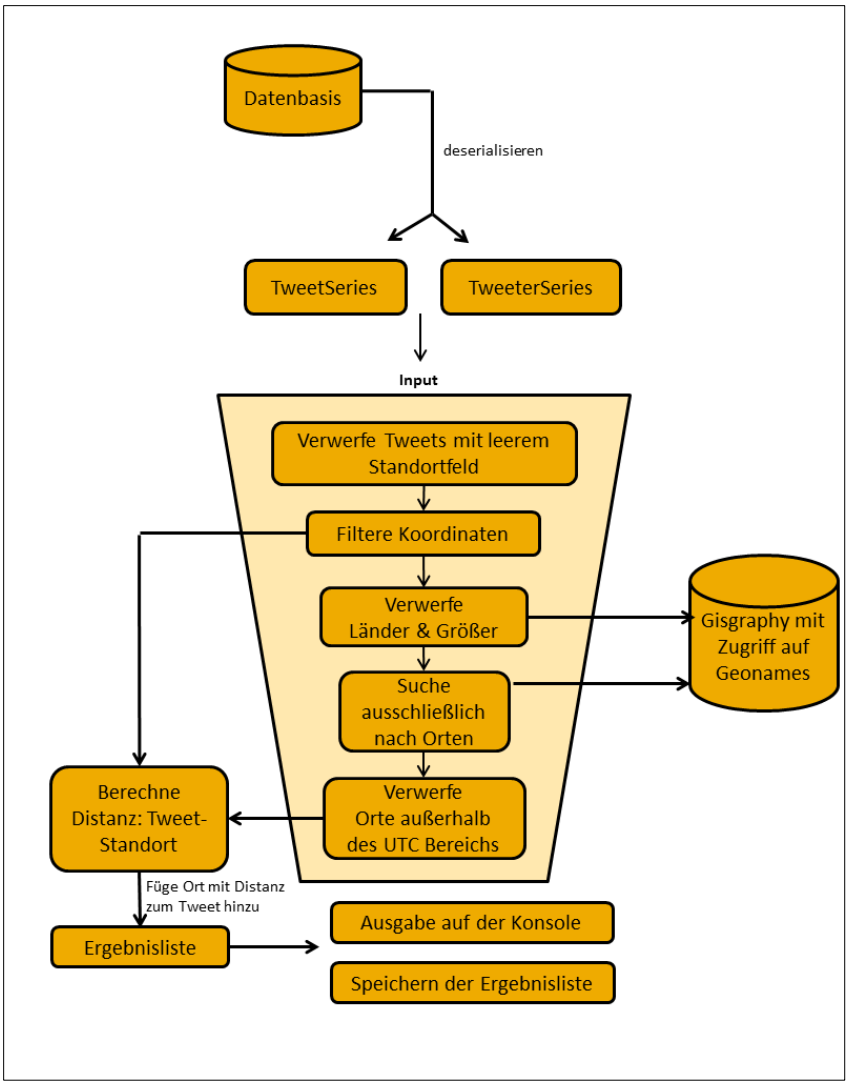


Abbildung 25: Ablaufbeschreibung der Implementierung

[32] Vorstellung Geonames und Gisgraphy im Kapitel 2.4.2.

Die Ergebnisse werden in folgender Tabellenform ausgegeben bzw. gespeichert.

Tweet Id	Best Nr.	Bester Ortsname	Beste Distanz	Durchschnitt über alle Orte	Koordinaten des Tweets	Tweeter Id	Inhalt des Standortfelds	**Ergebnisliste**

Ergebnisliste: enthält alle Rückgabewerte der Geonamesanfrage (Zeichenkette vom Standortfeld als Suchbegriff)						
Gefundener Ort	Koordinaten	Distanz	**Gefundener Ort**	Koordinaten	Distanz	…

In der Spalte **Bester Orts Name** werden der zum Standort[33] nächste Ort aus der Ergebnisliste gespeichert und in **Beste Distanz** deren Entfernung zueinander.

Statistikanalyse

Für die Analyse der statistischen Verteilung der Entfernung zwischen der Position des Tweets und dem Standort werden die zuvor gespeicherten Daten in Excel geladen. Anschließend wird eine Auswertung mit den folgenden statistischen Kennzahlen vorgenommen:

Ermittlung des Recalls der Geonamessuche
Gibt die relative Häufigkeit der Treffer der Geonamessuche zu den Suchanfragen an.

Geonamessuche: $\frac{\sum gefundener\ Orte}{\sum Suchanfragen}$

Ermittlung des Recalls der Tweetlokalisierung
Stellt die relative Häufigkeit von lokalisierten Tweets zur Gesamtmenge der Tweets dar. Es erfolgt keine Prüfung, ob der Tweet richtig verortet wurde.

Tweetlokalisierung: $\frac{\sum lokalisierte\ Tweets}{Gesamtanzahl\ aller\ Tweets}$

Ermittlung des Medians
Die Berechnung des Medians erfolgt einzeln für jede Spalte mit Distanzwert.

Ermittlung des Mittelwerts
Ermittlung des Mittelwerts für jede einzelne Spalte der Distanz.

Histogramm
Aus den Entfernungen wird ein Histogramm erstellt.
Die Einteilung der Distanzen in Kilometern erfolgt mit folgenden Klassen: 1, 2, 5, 10, 25, 50, 100, 1000, größer als 1000. Im Histogramm sind die relative und die kumulierte Häufigkeit der Klassen angegeben.

[33] Hier ist der Ort aus dem Standortfeld gemeint. Erläuterung im Kap. 1.2.

Anschließend erfolgt eine *Ausreißeranalyse*. Hierbei wird betrachtet, weshalb einige der Werte über dem Distanzwert von 60 km liegen. Die Intension ist es Optimierungsmöglichkeiten zu erkennen. Für Ausreißer, welche sich in verschiedene Klassen einordnen lassen, erfolgt eine Mehrfachzuordnung. Folgende Klassenbildung hat sich dabei etabliert (Tabelle 30):

Klasse	**Beschreibung**
Fehler Geonames	Es wird nicht der (richtige) Ort gefunden, welcher im Standortfeld eingetragen ist. Beispiel: Statt der gesuchten Stadt „*Rome*“ wird die Straße „*Rome 10, 7890 Ellezelles, Belgien*“ gefunden.
Länder & Kontinente	Im Standortfeld sind Bezeichner für Länder oder Kontinente enthalten.
Bundesländer & Bezirke	Eintragen von Bundesländern oder Bezirken im Standortfeld.
Mehrere Orte	Mehrere Eintragungen von Ortsnamen existieren.
Kein geografischer Bezug	Die Zeichenkette des Standortfelds enthält keine geografischen Bezüge.
Nicht am Standort	Der Tweeter befindet sich nicht am Standort.
Abkürzung	Das Standortfeld enthält Akronyme, welche nicht richtig zugeordnet werden können.
Keine Aussage	Zum Ausreißer lässt sich keine Ursache ermitteln.

Tabelle 30: Die Klassen der Ausreißeranalyse

Mit der *Null-Analyse* wird die Ursache eruiert, weshalb zu den Standortfeldangaben kein Eintrag in Geonames gefunden wurde (Tabelle 31).

Klasse	**Beschreibung**
Kein geografischer Bezug	Das Standortfeld enthält keine Angabe mit geografischem Bezug.
Fehler Geonames	Geonames enthält nicht den gesuchten Ort.
Keine Aussage möglich	Die Ursache für den fehlenden Eintrag kann nicht ermittelt werden.
Mehrere Orte	Mehrere Orte sind als Eintragungen vermerkt.
Rechtschreibfehler/ Schreibweise	Die Ortsangabe enthielt Rechtschreibfehler oder ist in einer unüblichen Schreibweise geschrieben.
Abkürzungen	Das Standortfeld enthielt nicht genormte[34] geografische Abkürzungen.

Tabelle 31: Die Klassen der Null-Analyse

[34] Nicht ISO 3166 entsprechende Einträge.

5. Evaluation des Ansatzes

In diesem Abschnitt werden die Ansätze aus Kapitel 4 evaluiert und die daraus resultierenden Ergebnisse miteinander verglichen. An erster Stelle wird die Referenz (Baseline) beschrieben, anschließend wird die Analyse der einzelnen Optimierungen durchgeführt. Dabei wird untersucht, wie sich die Suche nach ausschließlich Städten oder die Filterung von großflächigen Arealen[35] auf die Ergebnisse auswirken.
Zur Beurteilung der einzelnen Ergebnisse werden das statistische Lagemaß Median[36] und das arithmetische Mittel verwendet. Um die Entfernungen zwischen Tweet und Standort aus dem Nutzerprofil zu verdeutlichen, präsentiert die Grafik 26 die Stadt Darmstadt mit den eingezeichneten Radien im Abstand von ein (rot), zwei (grün) und fünf (blau) Kilometern. Zu beachten ist, dass größere Städte und Metropolen, wie beispielsweise Berlin, eine Ausdehnung von 20 km Radius und größer haben können[37].

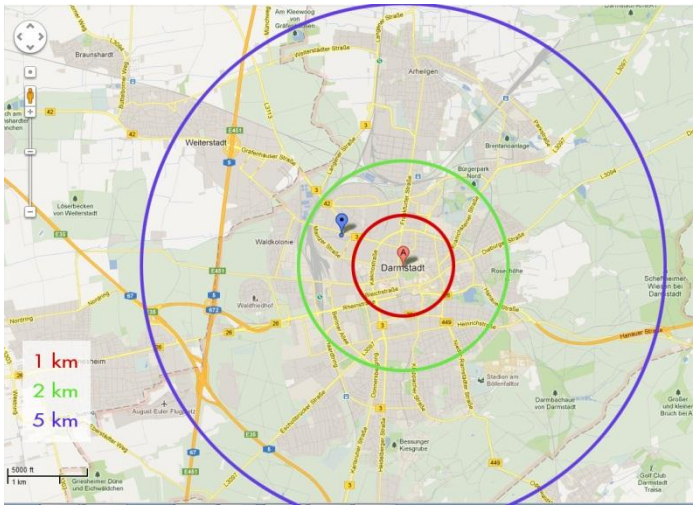


Abbildung 26: Googlemaps von Darmstadt mit eingezeichneten Radien

5.1. Beschreibung der Datenbasis

Für die Datenbasis wurden 168.409 Tweets mit den dazugehörigen Informationen des Tweeters verwendet. Die genutzten Profilinformationen umfassen das UTC-Offset und die Zeichenkette des Standortfelds. Die Daten wurden mit Streaming API von Twitter aufgezeichnet und umfassen den Zeitraum vom 15.12.2011 bis 24.12.2011. Jeder Tweet ist mit einem GPS-Koordinatenpaar versehen, um die Entfernung zur prognostizierten Position messen zu können. In 14,97 % der Fälle ist das Standortfeld nicht gesetzt. Insgesamt sind somit 143.204 Tweets mit Eintrag im Standortfeld in der Datenbasis existent. Diese Tweets sind in Tabelle 32 als valide Tweets bezeichnet. Insgesamt befinden sich 10.135 Koordinatenpaare in dieser Menge.

Tweets insgesamt	168.409
Mit leerem Standortfeld	25.205 (14,97 %)
Valide Tweets insgesamt	**143.204 (85,03 %)**
Koordinatenpaare im Standortfeld[38]	10.135
Valide Tweets ohne Koordinaten	**133.069**

Tabelle 32: Beschreibung der Datenbasis

[35] Erläutert in Kapitel 4.2.2. Gemeint sind Länder, Kontinente und Ozeane.
[36] Der Median ist der Wert in der Mitte einer sortierten Liste und ist robuster gegenüber Ausreißern.
[37] Vgl. Brockhaus (93) Band 23, S. 551.
[38] Koordinaten vom Typ Dezimalgrad. Beispiel 48.156509, 11.502149.

5.2. Baseline

Die Baseline stellt eine Referenz für den Vergleich der weiteren Optimierungen dar. Für diesen ersten Versuch erfolgten keine Einschränkungen des Ortstyps und keine Filterung bei der Suche mit Gisgraphy. Für die Experimente werden alle Koordinateneintragungen des Standortfelds zuvor entfernt, um eine unverfälschte Betrachtung der Ergebnisse zu ermöglichen, da die eingesetzten Filter keine Koordinatenangaben selektieren würden. Sinkt der Recall durch die Filterung und bleibt die Anzahl der Koordinatenpaare in der Ergebnismenge gleich, tritt eine Verzerrung in Richtung der Ergebnisqualität der Koordinaten auf.
Insgesamt stehen somit 133.069 Tweets ohne Koordinatenangaben zur Verfügung. Die Betrachtung der Tweets mit Koordinateneintragungen erfolgt separat im Abschnitt 5.3.4.

Für 85.379 von 143.022 Tweets mit validen Standortangaben wurde ein Ort gefunden, dies entspricht einer Quote von 64,16 %. Könnte stets der beste Ort aus der Ergebnisliste gewählt werden, so beträgt der Median 9,139 km - im Vergleich dazu der Median der ersten Listenposition mit 18,304 km. Damit wird deutlich, wie groß der Einfluss der Ortswahl aus der Ergebnisliste auf den Median ist. Der Tabelle 33 ist die statistische Auswertung der Ergebnisse der Baseline zu entnehmen.

Anzahl lokalisierter Tweets	85.379
Recall Geonames	64,16 %
Recall Geonames zur Baseline	100 %
Recall insgesamt verorteter Tweets	50,70 %
Median beste Listenposition	9,139 km
Median erste Listenposition	18,304 km
Median zweite Listenposition	52,136 km
Arith. Mittel beste Listenposition	701 km
Arith. Mittel erste Listenposition	1.329 km
Arith. Mittel zweite Listenposition	2.208 km

Tabelle 33: Ergebnisse der Baseline

Die Tabelle 34 beschreibt die Verteilungen der Entfernung bei der Wahl des Orts mit der geringsten Distanz zum Tweet aus der Ergebnisliste der Geonamesabfrage. Damit ist verdeutlicht, was maximal an Genauigkeit für den Fall möglich ist, dass stets die beste Position aus der Ergebnisliste gewählt werden könnte. Das Diagramm 27 präsentiert diese Werte. Auf der X-Achse ist der Radius der Entfernung zwischen Tweet und dem Standort aus dem Nutzerprofil angeben. Die linke Y-Achse gibt die relative Häufigkeit (blauer Balken) der Tweets im jeweiligen Radius an, die rechte Y-Achse stellt die kumulierte dar (rote Linie mit den Prozentangaben).

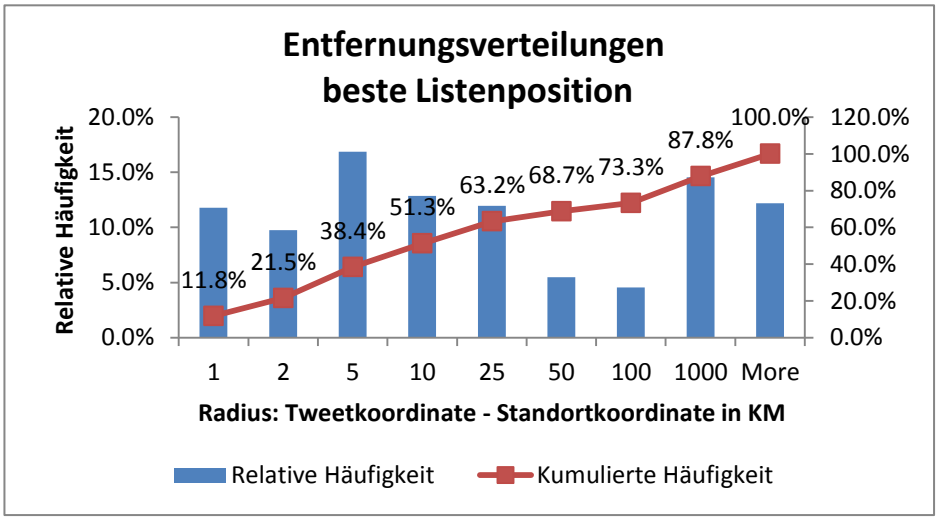


Abbildung 27: Diagramm der Entfernungsverteilungen vom Tweet zum besten Ort

Die Abbildung 28 visualisiert die Wahrscheinlichkeitswerte der Tabelle 34, dass sich ein Tweet im Umkreis um den prognostizierten Ort befindet.

Entfernung in Kilometern	Absolute Häufigkeit	Relative Häufigkeit	Kumulierte Häufigkeit
1	10.060	11,8 %	11,8 %
2	8.317	9,7 %	21,5 %
5	14.401	16,9 %	38,4 %
10	10.988	12,9 %	51,3 %
25	10.219	12,0 %	63,2 %
50	4.688	5,5 %	68,7 %
100	3.889	4,6 %	73,3 %
1000	12.418	14,5 %	87,8 %
>1000	10.399	12,2 %	100,0 %

Tabelle 34: Distanzverteilung zwischen Tweet und bestmöglichem Ort aus der Ergebnisliste

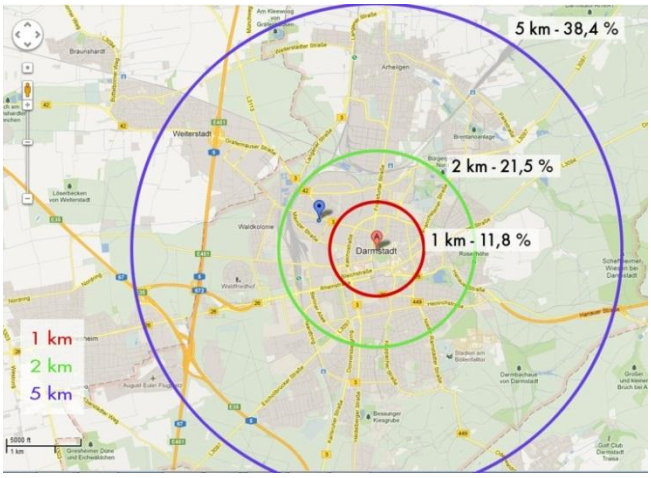


Abbildung 28: Googlemaps von Darmstadt mit eingezeichneten Distanzradien und Wahrscheinlichkeit, dass sich ein Tweet in deren Bereich befindet.

Bei der Analyse der Ausreißer wurden 101 Tweets mit einer Distanz von über 60 Kilometern zum prognostizierten Standort untersucht. Bei der Klassifikation fand eine Mehrfachzuordnung für die Ausreißer statt, welche sich gleichzeitig in verschiedene Klassen einordnen lassen. Die Tabelle 35 stellt die Ursachen der Ausreißer und deren Häufigkeit dar. Am weitesten verbreitet ist, dass der Tweeter die Nachricht nicht am eingetragenen Standort schreibt. Weiterhin kommen Eintragungen von größeren geografischen Arealen, wie Kontinenten, Ländern und Bundesländern im Standortfeld vor, hier als Land deklariert. Die Abweichung erklärt sich dadurch, dass größere Areale nur mit einem Koordinatenpaar angeben sind. Es lässt sich keine Aussage treffen, ob die Twitternachricht in dem angebenden Areal liegt.

Klasse	Absolute Häufigkeit
Nicht am Standort	48
Land	23
Kein geografischer Bezug	17
Fehler Gisgraphy/Geonames	11
Bundesland	5
Abkürzung	4
Mehrere Orte	1
Keine Aussage	1

Tabelle 35: Ergebnisse der Ausreißeranalyse der Baseline

Beispiele für gefundene Einträge zu Standortfeldern ohne vermuteten geografischen Bezug:

- *„Couch"* mit *Township of Couch*
- *„Sirius 6B[39]"* mit Mount Sirius
- *„Krypton[40]"* als Ort in den USA

[39] „Sirius 6B" ist ein Planet aus einem Science-Fiction Film.

[40] Krypton ist ein Planet aus einem Comic.

In der Tabelle 36 erfolgt die Vorstellung der vermuteten fehlerhaften Zuordnung von Geonames.

Eintragung im Standortfeld	Gefundener Ort	Bemerkung
„0320“	Rustenburg	*0320* ist eine Postleitzahl aus der Province Rustenburg aus Süd Afrika.
„UK“	Pangch‘ uk-kol	Die nordkoreanische Stadt enthält *uk* im Namen. Als Alternative wird beispielweise die Straße „UK“ in Russland vorgeschlagen.
„A T L A N T A“	Tân Phú, Châu Thành (Vietnam)	Es liegt kein Wort vor, nur einzelne Buchstaben.
„Santiago“	Santiago de Compostela	Mehrere Santiagos sind existent. Herausforderung der Ortswahl.
„Saudi Arabia - Al-Qassim“	Saudia Arabia	Aufgrund der Schreibweise „Land - Ort“ findet Gisgraphy nur das Land.

Tabelle 36: Analyse der fehlerhaften Zuordnung von Gisgraphy/Geonames

Die Analyse der Eintragungen aus dem Standortfeld ohne Ergebnis bei der Abfrage von Geonames ergibt die in Tabelle 37 aufgelisteten Ergebnisse. Untersucht wurden 50 dieser Fälle. Bei der genaueren Betrachtung ist festzustellen, dass die Volltextsuche keine Ergebnisse liefert, wenn Rechtschreibfehler vorliegen, der Ort in einem Satz steht und wenn mehrere Orte vorkommen. Die Analyse ergab weiterhin, dass die genaue Punktierung bei Kommagruppen wichtig ist. Der Ort „lawrenceville , Ga“ wird aufgrund der Kommafehlstellung nicht gefunden.

Klasse	Häufigkeit	Beispiel
Kein geografischer Bezug	24	*„The Land Below the Wind“*
Fehler Gisgraphy/Geonames	7	*„Bosten Masss“, „lawrenceville , Ga“*
Keine Aussage möglich	7	*„La o tu n'es pas .“*
Zwei Orte	6	*„Melbourne-Riyadh“, „Munich / Tuebingen, Germany“*
Rechtschreibfehler/ Schreibweise	4	*„barcelon**aa**“, „I´m in MIAMI bit[…]!”*
Abkürzungen	3	*„D[M]V“*

Tabelle 37: Null-Analyse der Baseline

5.3. Evaluation der Filterung

In diesem Abschnitt werden die Filterungen und deren Kombinationen, welche im Kapitel 4.2.2. vorgestellt worden sind, evaluiert. Wie zuvor bei der Baseline erläutert, werden die Koordinaten-angaben im Standortfeld für die kommenden Versuche nicht betrachtet.

5.3.1. Experiment: Einschränkung der Suche auf Städte

Beim ersten Experiment wird ausschließlich nach Städten (Ortstyp: City) im geografischen Lexikon Geonames gesucht. Die Tabelle 38 präsentiert die Ergebnisse des Experiments. Im Diagramm 29 ist Verteilung der Distanzen dargestellt.
Im Vergleich zur Baseline ist der Median der ersten Position in der Ergebnisliste um 2,25 km auf 16,05 km gesunken. Gleichzeitig ist der Wert der besten Listenposition von 9,24 km auf 10,15 km gestiegen. Dieser Effekt lässt sich folgendermaßen erklären: Bei der Baseline wurden auch Lokalitäten innerhalb der Stadt gefunden, wie beispielsweise Hotels. Erfolgt die Suche eingeschränkt auf Städte, so wird nur die Stadt als Ergebnis zurückerhalten und keine Örtlichkeiten innerhalb dieser.
Wie zu erwarten war, ist die Anzahl der zu verortenden Tweets auf 94,4 % im Vergleich zur Baseline gesunken. Der Mittelwert ist stark angestiegen. Dies deutet darauf hin, dass die Ausreißer in Anzahl und Relevanz angestiegen sind. Die Ermittlung der Ursache für die Verdopplung des Mittelwerts erfolgt in der Ausreißeranalyse.

Anzahl der Ergebnisse	80.920
Recall Geonames	60,80 %
Recall Geonames zur Baseline	94,40 %
Recall insgesamt verorteter Tweets	48,05 %
Median beste Listenposition	10,15 km
Median erste Listenposition	16,05 km
Median zweite Listenposition	560,55 km
Arith. Mittel beste Listenposition	794 km
Arith. Mittel erste Listenposition	2.834 km
Arith. Mittel zweite Listenposition	2.927 km

Tabelle 38: Experiment: Einschränkung der Suche auf Städte

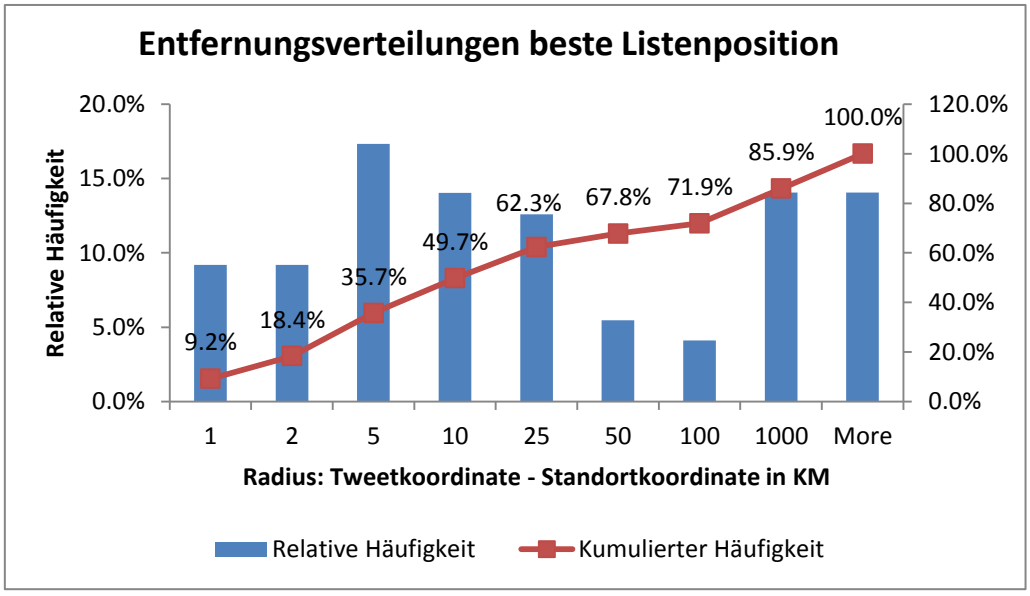


Abbildung 29: Diagramm der Entfernungsverteilungen vom Tweet zum besten Ort

Ausreißeranalyse:

Es wurden 100 Ausreißer mit einer Abweichung von über 60 km untersucht. Tabelle 39 veranschaulicht die Ergebnisse. Die häufigste Ursache ist, dass sich der Tweeter nicht am Standort befindet. Im zweithäufigsten Fall lag die Herausforderung in der Interpretation der Ortsangabe von Gisgraphy sowie in der Auswahl des Orts aus der Ergebnisliste, in der Tabelle 39 als Fehler Geonames bezeichnet.
Zu beachten ist, dass alle Länderangaben bei der Suche nach ausschließlich Städten als Stadt interpretiert werden. Dies beinhaltet ein großes Fehlerpotenzial. In der Tabelle 40 werden einige Ausreißer intensiver untersucht.

Klasse	**Absolute Häufigkeit**
Fehler Geonames	27
Länder & Kontinente	16
Bundesländer & Bezirke	9
Mehrere Orte	1
Kein geografischer Bezug	11
Nicht am Standort	30
Abkürzung	3
Keine Aussage	1

Tabelle 39: Ergebnisse der Ausreißeranalyse

Eintragung im Standortfeld	**Gefundener Ort**	**Bemerkung**
„São Paulo – SP“	São Paulo do Potengi	Das Leerzeichen vor dem Bindestrich verursacht die Fehlverortung nach São Paulo do Potengi.
„Belgium“	Belgium, Ort in Illinois	Vermutlich ist mit dem Eintrag im Standortfeld das Land gemeint.
„SBC“	São Bernardo do Campo	Das Akronym wird als Stadt in Brasilien verortet.
„Global“	Global Village, Ort in Indien	Vermutlich hat der Eintrag keinen geografisch gemeinten Bezug.
„Dreamland'“	Dreamland Mobile Home Park als Stadt in den US verzeichnet.	
„SPRING ,TX“	Cherry Spring	In Geonames befindet sich nicht das vermutlich gemeinte *Spring* an erster Position.

Tabelle 40: Analyse einzelner Ausreißer

5.3.2. Experiment: Filterung von Ländern & größeren Arealen

In diesem Experiment werden alle Länder und größeren Areale, darunter fallen Kontinente und Ozeane, vor der Suche gefiltert. Die Intension ist die Vermutung, dass die Genauigkeit (Median und Mittelwert) präziser wird, die Anzahl (Recall) der zu verortbaren Tweets sinkt. Zur Filterung wird mit Gisgraphy eine Anfrage gestellt, ob der Suchbegriff ein Land ist, und bei positivem Ergebnis entfällt die Lokalisierung des Tweets. In der Tabelle 41 sind die Ergebnisse der Filterung verzeichnet und Abbildung 30 visualisiert die Entfernungsverteilungen. Die Filterung verortet insgesamt weniger Tweets im Vergleich zu den vorherigen Experimenten. Der Median der besten Position sowie derjenige der ersten Position der Rückgabeliste sind deutlich genauer geworden. Im Vergleich zur Baseline ist der Median der ersten Position vom 18,30 km auf 13,62 km gesunken. Bei der Ergebnisanalyse wurde festgestellt, dass der Filter alle Länder zuverlässig verworfen hat. Im nächsten Abschnitt erfolgt eine detaillierte Betrachtung des Filters.

Anzahl lokalisierter Tweets	75.699
Recall Geonames	56,88 %
Recall Geonames zur Baseline	88,31 %
Recall insgesamt verorteter Tweets	44,94 %
Median beste Listenposition	7,40 km
Median erste Listenposition	13,62 km
Median zweite Listenposition	33,45 km
Arith. Mittel beste Listenposition	667 km
Arith. Mittel erste Listenposition	1.211 km
Arith. Mittel zweite Listenposition	1.897 km

Tabelle 41: Experiment: Filterung der Länder und Regionen

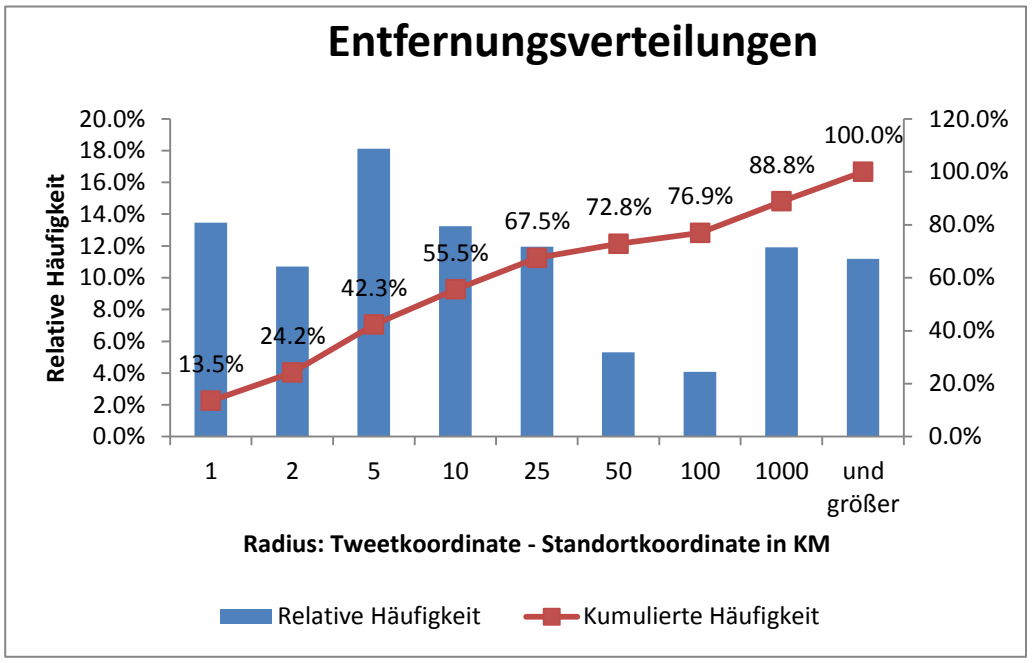


Abbildung 30: Diagramm der Entfernungsverteilungen vom Tweet zum besten Ort

Evaluation des Filters *Länder & größere Areale*

In diesem Abschnitt erfolgt eine Evaluation des *Länder & größere Areale* Filters. Dabei wurden 1000 zufällig gewählte Tweets verwendet. Von diesen befinden sich 143 ohne Angaben im Standortfeld und werden nicht betrachtet. Bei 59 wurde ein Land[41] gefunden, das entspricht knapp 6,9 %. Es erfolgte die Messung der Distanz zwischen den Koordinaten des Tweets und denjenigen des Lands. Von den 59 Ländereintragungen haben 46 eine Distanz von über 60 km (Tabelle 42). Der Median der ersten Position der Liste beträgt 285,28 km und ist somit über das 15-Fache höher.

Die manuelle Betrachtung hat ergeben, dass Abkürzungen als Land klassifiziert werden können. Beispielsweise ist zu *„UK“* das United Kingdom als Treffer verzeichnet.

Median erste Position	285,28 km
Mittelwert	2.062 km
Häufigkeit der Distanzen größer 60 km	46 (80 %)
Anzahl der fehlerhaft Klassifizierten	4 (6,8 %)

Tabelle 42: Ergebnisse der Filterevaluation

In der Tabelle 43 sind die Beispiele der Klassifikation aufgezeigt, für welche die Suche mit Gisgraphy fehlschlagen kann. So sind Angaben in Form von *„Land, Stadt“* oftmals als Land verortet. Weiterführende Untersuchungen haben nachgewiesen, dass diese Fehlverortung nur in einigen Fällen stattfindet. Bei „Deutschland, Weiterstadt“ wird die Stadt Weiterstadt gefunden.
Weiterhin werden alle Bezeichner, welche gleichzeitig einen Ländername und einen Ortsname darstellen, gefiltert. Beispielsweise existieren mehrere Orte mit dem Namen St. Vincent, jedoch auch ein Inselstaat in der Karibik mit diesem Namen.

Eintrag Standortfeld	**Gefundener Eintrag**	**Ursache**
„Guinea, Conakry”	Papua New Guinea	Landangabe vor Stadtangabe
„NY“	Papua New Guinea	*NY* ist ähnlich zum Länder Code von Papua New Guinea *NC*
„Salvador - Bahia – Brasil“	El Salvador	Ähnlich klingender Ländername
„Saudi Arabia – Riyadh“	Saudi Arabia	Landangabe vor Stadtangabe

Tabelle 43: Analyse der fehlerhaften Klassifikation

[41] Länder & größere Areale - im Folgenden als Land bezeichnet.

5.3.3. Experiment: Filterung von Ländern & größeren Arealen und Einschränkung der Suche auf Städte

In diesem Abschnitt erfolgen die Vorstellung des Experiments aus der Kombination des *Länder & größere Areale* Filters und die Einschränkung der Suche auf ausschließlich Städte. Die Tabelle 44 und Abbildung 31 stellen die Ergebnisse vor. Beide Filter stellen sich als wirksam heraus. Der Median der ersten Position ist um 1,52 km gegenüber der Filterung von nur Ländern & größeren Arealen gesunken. Der leichte Anstieg des Medians der bestmöglichen Position sowie der des Mittelwerts der ersten Position erklären sich aus der Sucheinschränkung auf Städte. Der Effekt wurde in Kapitel 5.3.1. beschrieben.

Anzahl lokalisierter Tweets	70.905
Recall Geonames	53,28 %
Recall Geonames zur Baseline	82,72 %
Recall insgesamt verorteter Tweets	42,10 %
Median beste Listenposition	8,111 km
Median erste Listenposition	12,099 km
Median zweite Listenposition	370,765 km
Arith. Mittel beste Listenposition	660 km
Arith. Mittel erste Listenposition	1.249 km
Arith. Mittel zweite Listenposition	2.667 km

Tabelle 44: Experiment: Filterung von Ländern & Co und Einschränkung der Suche auf Städte

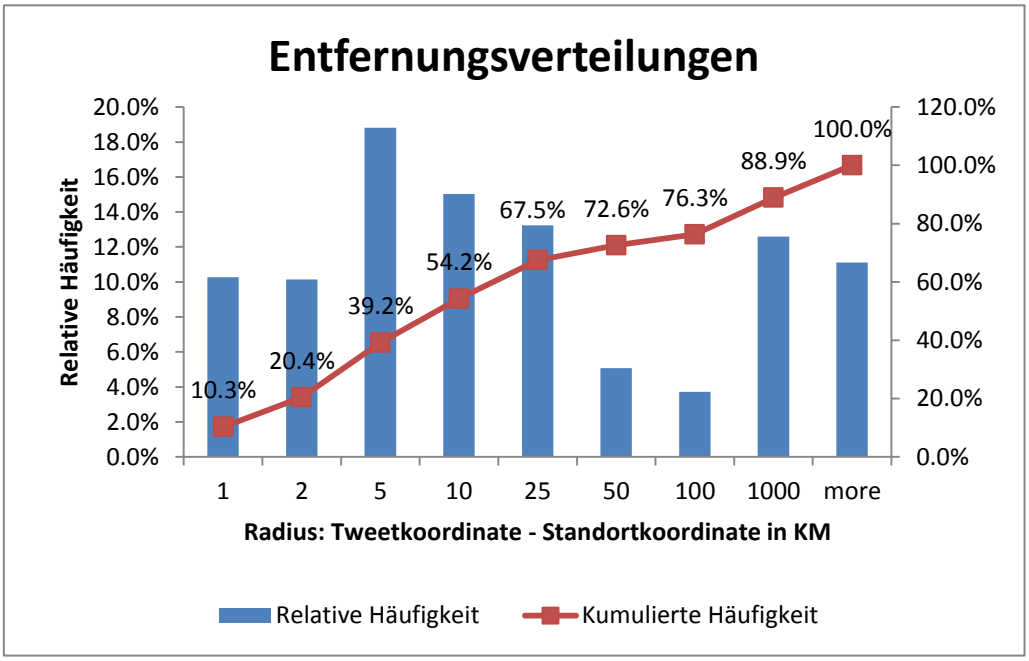


Abbildung 31: Diagramm der Entfernungsverteilungen vom Tweet zum besten Ort

Ausreißeranalyse

Dieser Analyse von 100 Ausreißern (Tabelle 45) ist zu entnehmen, dass der Filter *Länder & größere Areale* alle Länder herausgefiltert hat. Mit 53 Einträgen ist die häufigste Ursache für die Prognoseabweichung von größer 60 km die Tatsache, dass der Nutzer sich nicht am Standort befindet. Die Tabelle 46 stellt benennt Beispiele für Ausreißer und Herausforderungen.

Klasse	**Beschreibung**
Fehler Geonames	20
Länder & Kontinente	0
Bundesländer & Bezirke	15
Mehrere Orte	4
Kein geografischer Bezug	7
Nicht am Standort	53
Abkürzung	1
Keine Aussage	2

Tabelle 45: Ergebnisse der Ausreißeranalyse

Eintragung im Standortfeld	**Gefundener Ort**	**Bemerkung**
„NRW“	Lembeck	Lembeck liegt in Nordrhein-Westfalen und wird vermutlich gefunden, weil sich im Link von Wikipedia die Zeichenfolge NRW befindet. en.wikipedia.org/wiki/Lembeck_%28**NRW**
„Mars“	Saint-Mars-la-Jaille	Vermutlich hat die Standortfeldangabe „Mars“ den Bezug zum Planeten „Mars“.
„Hampshire“	Hampshire	Zum Bezeichner *„Hampshire“* existierende Orte.
„Milford, DE“	Orte im Bundesstaat Delaware (USA)	An diesem Beispiel wird die Ambiguität der Akronyme deutlich. DE kann sowohl Deutschland als auch Delaware abkürzen.
„Newark, DE“		

Tabelle 46: Analyse der fehlerhaften Klassifikation

5.3.4. Experiment: Einschränkung auf Koordinaten

Bei diesem Versuch wurde die Distanz zwischen dem Tweet und dem im Standortfeld eingetragenen Koordinatenpaar ermittelt. Insgesamt sind 10.133 Koordinatenpaare enthalten, das entspricht 6,02 % des Datensets. In einem Umkreis von 3,388 km um die Prognose liegen 50 % aller Tweets. Das Diagramm 32 veranschaulicht die Verteilung der Tweets über die Distanzen.

Die manuelle Untersuchung der Koordinaten hat ergeben, dass vermutlich die Angabe *„0.0, 0.0“* fehlerhaft ist. Diese liegt im Golf von Guinea. Üblicherweise sind Koordinaten mit mehreren Nachkommastellen angeben und es ist als unwahrscheinlich anzunehmen, dass diese Tweets alle von exakt dieser GPS-Position stammen.

Anzahl der Ergebnisse	10.133
Recall verortete Tweets	6,02 %
Recall verortete Tweets zur Baseline	11,82 %
Fehlerhafte Koordinaten *„0.0, 0.0“*	131
Median	3,388 km
Mittelwert	515 km

Tabelle 47: Ergebnisse des Experiments: Betrachtung von ausschließlich Koordinaten

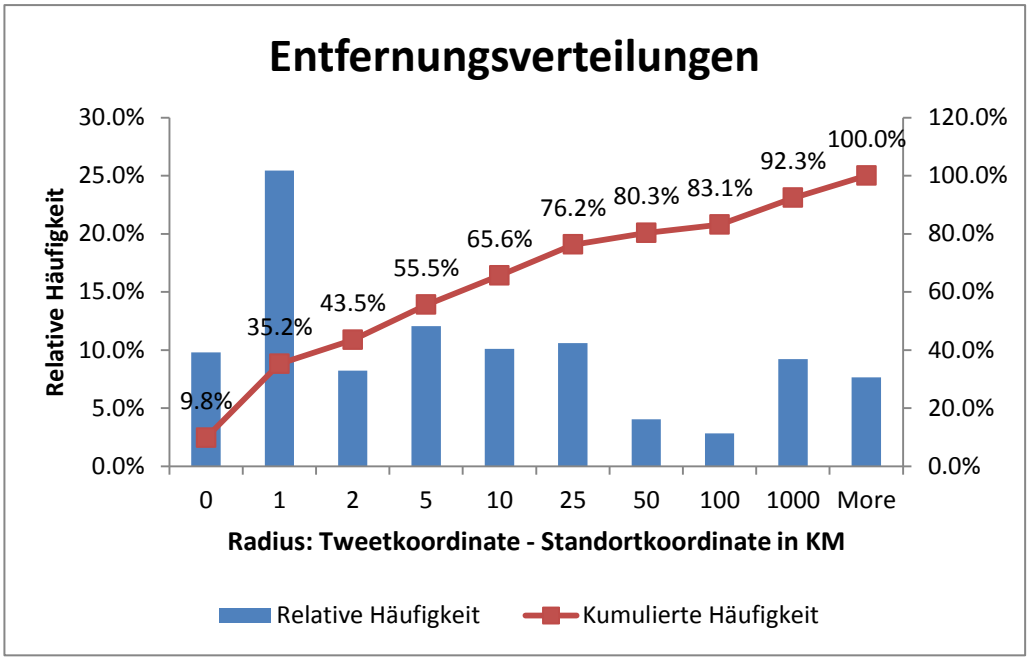


Abbildung 32: Diagramm der Entfernungsverteilungen vom Tweet zur Koordinate aus dem Standortfeld

Abbildung 33 visualisiert die Wahrscheinlichkeitswerte der Tabelle 48 dahingehend, dass sich ein Tweet im Umkreis um den prognostizierten Ort befindet. Es befinden sich 35,2 % der Tweets im Radius von einem km, 43,5 % in zwei km und 55,5 % in fünf km um die prognostizierte Position.

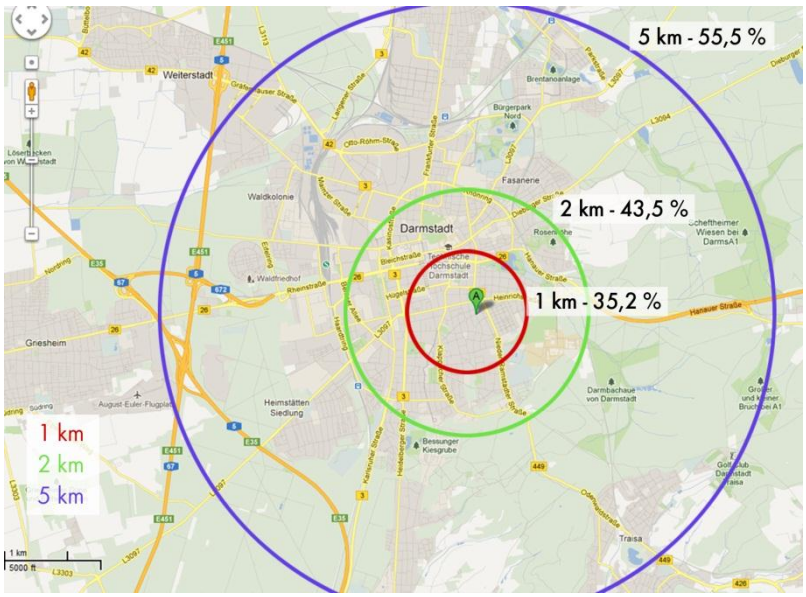


Abbildung 33: Googlemaps von Darmstadt mit eingezeichneten Distanzradien und Wahrscheinlichkeit, dass sich ein Tweet in deren Bereich befindet.

Entfernung in Kilometern	Häufigkeit	Relative Häufigkeit	Kumulierte Häufigkeit
0	993	9,8 %	9,8 %
1	2.578	25,4 %	35,2 %
2	833	8,2 %	43,5 %
5	1.222	12,1 %	55,5 %
10	1.025	10,1 %	65,6 %
25	1.075	10,6 %	76,2 %
50	409	4,0 %	80,3 %
100	288	2,8 %	83,1 %
1000	934	9,2 %	92,3 %
>1000	776	7,7 %	100,0 %

Tabelle 48: Entfernungsverteilungen

5.4. Auswahl eines Orts aus der Ergebnisliste

Bei der Suche mit Gisgraphy in dem geografischen Lexikon Geonames wird eine Liste mit Ergebnissen zurückerhalten. Es besteht die Notwendigkeit aus dieser Liste einen Ort auszuwählen. Im ersten Abschnitt erfolgt die Erarbeitung einer geeigneten Auswahlstrategie. Die einzelnen Strategien wurden im Kapitel 4.2.3. vorgestellt. Anschließend weist ein Experiment die Auswirkungen der Auswahlstrategie auf die Tweetlokalisierung nach.

5.4.1. Entwicklung einer geeigneten Auswahlstrategie

In diesem Abschnitt wird eine Kombination aus den Strategien zur Wahl eines Orts vorgenommen. Zuvor ist auf das Faktum zu verweisen, dass Gisgraphy die Ergebnisse nach einem Scorewert sortiert. Das genaue Verfahren ist nicht bekannt. Mit den vergangenen Experimenten wurde empirisch nachgewiesen, dass durchschnittlich die erste Listenposition den weiteren Positionen überlegen ist. Die Tabelle 49 listet die Ergebnisse der Baseline für die verschiedenen Positionen auf. Der Median der vorderen Ränge ist jeweils besser als bei den nachfolgenden.
Deshalb wird als Auswahlstrategie stets die erste Position gewählt.

Position in der Liste	**1**	**2**	**3**	**4**
Median	18,30 km	52,14 km	104,88 km	128,35 km
Mittelwert	1.329 km	2.208 km	2.181 km	2.351 km

Tabelle 49: Untersuchungen zur Ortsauswahlstrategie

Es besteht weiterhin die Herausforderung der Ambiguität der Ortsbezeichner. Als Beispiele dienen der Ort Darmstadt in Hessen und zwei gleichnamige Orte in den USA. Um eine Disambiguierung vorzunehmen, wird die UTC-Offset Eintragung aus dem Nutzerprofil verwendet. Nur diejenigen Orte, welche innerhalb des UTC-Bereichs des Nutzers liegen, stehen zur Auswahl.

Als Auswahlstrategie steht die erste Eintragung in der Ergebnisliste fest, welche innerhalb des UTC-Bereichs liegt.

5.4.2. Experiment: Auswahl der Orts mithilfe des UTC-Bereichs

In diesem Abschnitt erfolgt die Beurteilung der Güte des UTC-Bereichs bei der Lokalisierung der Tweets. Alle Orte außerhalb des UTC-Bereichs sind aus der Ergebnisliste herausgefiltert.

Das Resultat des Experiments lautet, dass insbesondere der Mittelwert der ersten Position mit 650 km im Vergleich zu den vorherigen Versuchen mit jeweils über 1200 km niedriger ist. Weiterhin wird auch der Median der Baseline hier unterboten. Zusammenfassend stellt sich die Wahl des Orts unter Zuhilfenahme des UTC-Bereichs als eine geeignete Methode heraus. Die Tabelle 50 präsentiert die detaillierten Ergebnisse und im Diagramm 34 ist die Entfernungsverteilung zwischen Tweet und dem Ort aus dem Standortfeld dargestellt.

Anzahl lokalisierter Tweets	57.907
Recall Geonames	43,52 %
Recall Geonames zur Baseline	67,56 %
Recall insgesamt verorteter Tweets	34,38 %
Median beste Listenposition	8,891 km
Median erste Listenposition	15,955 km
Median zweite Listenposition	25,092 km
Arith. Mittel beste Listenposition	566 km
Arith. Mittel erste Listenposition	650 km
Arith. Mittel zweite Listenposition	589 km

Tabelle 50: Experiment: Auswahl der Orts mithilfe des UTC-Bereichs

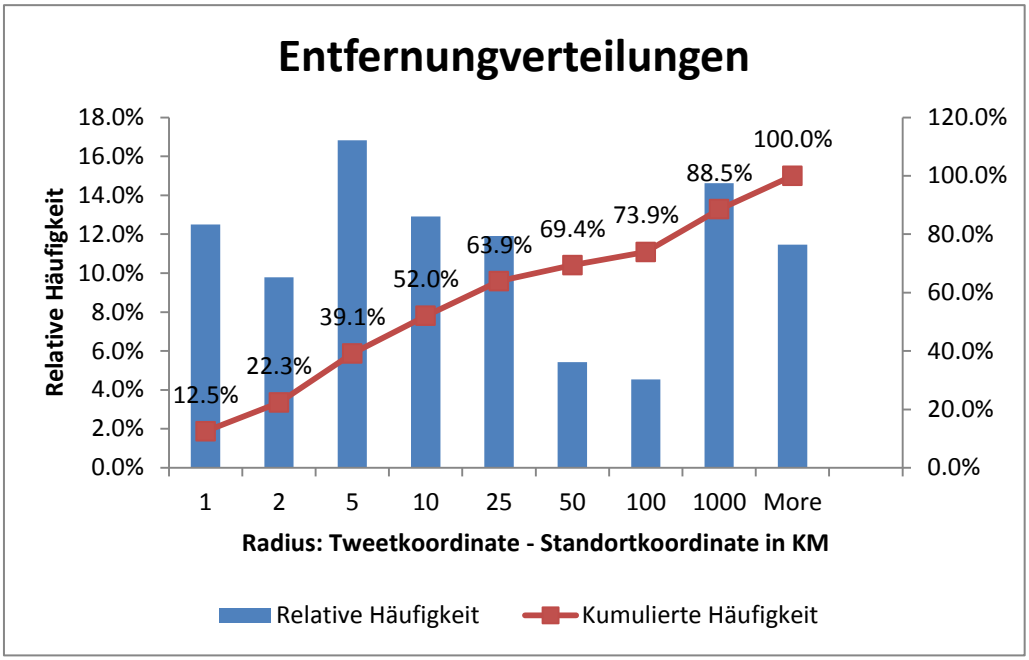


Abbildung 34: Diagramm der Entfernungsverteilungen vom Tweet zum besten Ort der Ergebnisliste

Ausreißeranalyse

Da keine Sucheinschränkung und keine Filterung vorgenommen werden, sind die Ausreißer im Bereich der Länder zu verzeichnen. Die Ergebnisse in der Tabelle 51 ähneln denen der vorangegangenen Untersuchungen. Die häufigste Ursache ist, dass der Tweeter sich nicht am Standort befindet.

Klasse	**Absolute Häufigkeit**
Fehler Geonames	19
Länder & Kontinente	22
Bundesländer & Bezirke	15
Mehrere Orte	2
Kein geografischer Bezug	3
Nicht am Standort	40
Abkürzung	2
Keine Aussage	3

Tabelle 51: Ergebnisse der Ausreißeranalyse

5.5. Experiment: Kombinationen der Filterung mit Verwendung des UTC-Bereichs

In diesem Abschnitt erfolgt die Kombination der zuvor evaluierten Filter in folgender Form:
Zuerst kommt es zum Einsatz des Filters *Länder & Größere Areale* und anschließend zur Sucheinschränkung auf Städte. Danach wird der erste Ort aus der Ergebnisliste gewählt, welcher sich im UTC-Bereich befindet.
Die Tabelle 52 präsentiert die Ergebnisse. Im Vergleich zur Baseline hat sich der Recall fast halbiert. Der Median der ersten Position ist hingegen von ursprünglich 18,30 km auf 10,73 km gesunken. Das bedeutet, es liegen durchschnittlich 50 % aller Tweets im Umkreis von 10,73 km um den prognostizierten Ort. Durch die Verwendung des UTC-Bereichs konnte das arithmetische Mittel auf 581 km gesenkt werden. Im Diagramm 35 ist die Verteilung grafisch dargestellt.

Anzahl lokalisierter Tweets	47.570
Recall Geonames	35,75 %
Recall Geonames zur Baseline	55,50 %
Recall insgesamt verorteter Tweets	28,27 %
Median beste Listenposition	7,723 km
Median erste Listenposition	**10,734 km**
Median zweite Listenposition	68,765 km
Arith. Mittel beste Listenposition	520 km
Arith. Mittel erste Listenposition	**581 km**
Arith. Mittel zweite Listenposition	747 km

Tabelle 52: Experiment: Kombination von Filterung und Ortsauswahl unter Verwendung des UTC-Bereichs

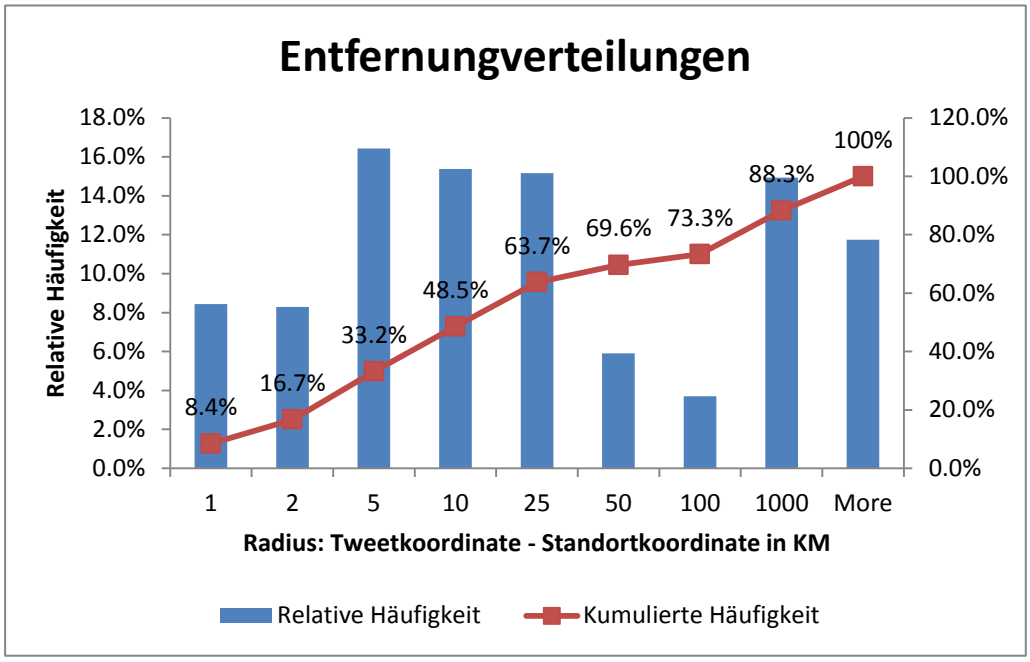


Abbildung 35: Diagramm der Entfernungsverteilungen vom Tweet zum ersten Ort der Ergebnisliste

5.6. Experiment durch Kombination der Filterung und Ortsauswahl unter Verwendung des UTC-Bereichs zzgl. Koordinaten

Das Experiment subsumiert die zuvor vorgestellte Kombination aus Kapitel 5.5. und fügt die zuvor filtrierten Koordinaten hinzu. Es ist bei weitem das präziseste Ergebnis, welches sich mit den vorgestellten Optionen erreichen lässt. Es lassen sich insgesamt 34,04 % aller Tweets lokalisieren. Die Hälfte von diesen liegt durchschnittlich im Umkreis von 9,233 km um die Prognose. Die Tabelle 53 präsentiert die Ergebnisse und das Diagramm 36 visualisiert die Verteilung der Tweets über die Distanzen.

Anzahl lokalisierter Tweets	57.331
Recall insgesamt verorteter Tweets	34,04 %
Median beste Listenposition	6,996 km
Median erste Listenposition	9,233 km
Median zweite Listenposition	68,77 km
Arith. Mittel beste Listenposition	517 km
Arith. Mittel erste Listenposition	567 km
Arith. Mittel zweite Listenposition	747 km

Tabelle 53: Experiment: Kombination der Filterung und Ortsauswahl unter Verwendung des UTC-Offsets zzgl. Koordinaten

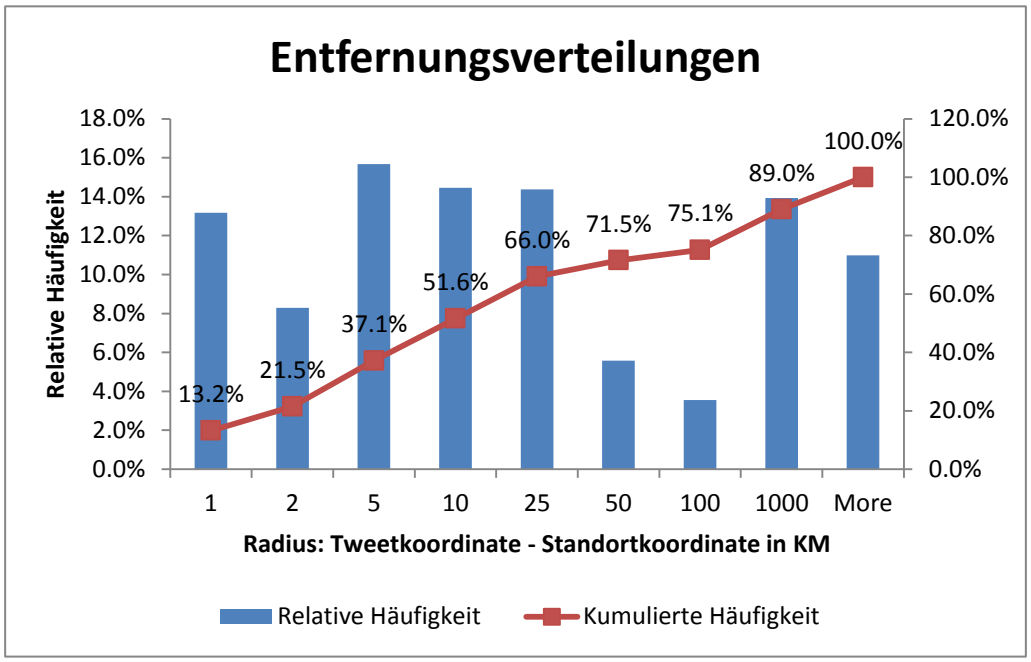


Abbildung 36: Diagramm der Entfernungsverteilungen vom Tweet zum ersten Ort der Ergebnisliste

5.7. Zusammenfassung und Fazit der Evaluation

In diesem Abschnitt werden die Ergebnisse der einzelnen Experimente zusammengefasst und ein Fazit der Evaluationen gezogen. Weiterhin wird aus den einzelnen Filtern und Ortsauswahloptionen ein Gesamtsystem zur Lokalisierung auf Basis des Standortfelds entwickelt.

Unter dem Gesamtsystem sind die Kombination der Filter sowie Optionen zu verstehen. Die Intention besteht darin, alle Tweets zu lokalisieren, die Verortungsgenauigkeit gegenüber der Baseline zu erhöhen und dem einzelnen Tweet die entsprechende Distanzverteilung zuzuordnen. Die Zuordnung der Lokalisierungsgenauigkeit zum Tweet funktioniert folgendermaßen: Der Eintrag im Standortfeld durchläuft die Filter und nach dem Prozess kann dem Tweet die Verteilungsfunktion des genauesten Filters zugeordnet werden. Die Verteilungsfunktionen der Filterungsmöglichkeiten wurden zuvor evaluiert und die Ergebnisse aller Experimente in der Tabelle 56 zusammengefasst.

Im Folgenden wird der Ablauf des Systems beschrieben, welcher in der Abbildung 37 verdeutlicht ist. Im ersten Schritt wird untersucht, ob sich im Standortfeld ein Koordinatenpaar befindet. Falls das zutrifft, kann dieses dem Tweet zugeordnet werden. Anderenfalls erfolgt eine Suche im geografischen Lexikon nach ausschließlich Städten. Wird kein Ergebnis zurückerhalten, so wird ohne Einschränkung wiederholt gesucht. Wird kein passender Ort gefunden, so kann der Tweet nicht lokalisiert werden.
Im Fall, dass bei den vorherigen Suchanfragen ein Ort zurückgegeben wurde, wird ein Abgleich des UTC-Bereichs mit dem gefundenen Ort und der Eintragung im Nutzerprofil vorgenommen. Ist nach der Filterung mit dem UTC-Bereich keine Lokalisierung möglich, so ist das Ergebnis der vorigen Suche zu verwenden.
Für die Orte, welche jedoch innerhalb des UTC-Bereichs liegen, erfolgt die Wahl des ersten Ortes aus der Liste. Am Ende des Systems liegen zu jedem Tweet eine Koordinate und die Wahrscheinlichkeitsverteilung vor, dass sich der Tweet im Umkreis der Prognose befindet.
Die Reihenfolge des Ablaufs ist irrelevant, solange immer das Ergebnis des präzisierten Filters gewählt wird. Die hier vorgestellte Reihenfolge optimiert die Rechenzeit, in dem die Abfragen an das geografische Lexikon minimiert werden.

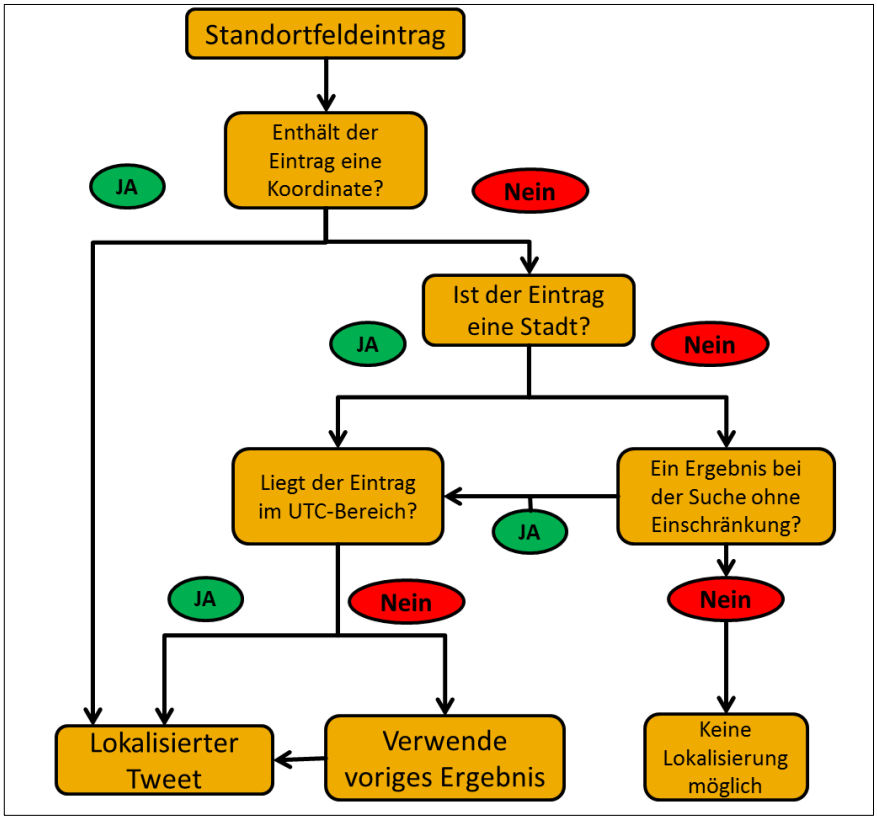


Abbildung 37: Ablaufbeschreibung des Systems

In der Tabelle 55 sind die Ergebnisse des zuvor beschriebenen Systems ohne Koordinaten aufgelistet. Damit lassen sich die Ergebnisse direkt mit der Baseline vergleichen. Der Median der ersten Position liegt mit 17,84 km nur marginal unter dem der Baseline mit 18,30 km und der Mittelwert ist um 102 km gestiegen.
Es lässt sich insgesamt nur eine geringfüge Verbesserung gegenüber der Baseline feststellen. Die Ursache liegt in der Tatsache begründet, dass keine Herausfilterung der Tweets stattfand, sondern dass mithilfe der Filter versucht wurde, für jeden Tweet den bestmöglichen Ort zu wählen.

Tabelle 54 fügt zu den Ergebnissen des Systems die zuvor herausgefilterten Koordinaten hinzu und repräsentiert das Endergebnis für alle lokalisierbaren Tweets. Damit lassen sich mit diesem Ansatz maximal 56,77 % aller Tweets mit einem Median von 15,22 km lokalisieren. Im Diagramm 38 ist die Verteilung der Distanzen zwischen der GPS-Position und dem prognostizierten Standort angegeben.

Anzahl lokalisierter Tweets	85.379
Recall Geonames	59,62 %
Recall Geonames zur Baseline	100 %
Recall insgesamt verorteter Tweets	50,70 %
Median beste Listenposition	12,04 km
Median erste Listenposition	**17,84 km**
Median zweite Listenposition	234,40 km
Arith. Mittel beste Listenposition	1.036 km
Arith. Mittel erste Listenposition	**1.431 km**
Arith. Mittel zweite Listenposition	1.897 km

Tabelle 55: Experiment: Ergebnisse des Systems ohne Koordinaten

Anzahl lokalisierter Tweets	95.604
Recall insgesamt verorteter Tweets	56,77 %
Median beste Listenposition	10,70 km
Median erste Listenposition	**15,22 km**
Arith. Mittel beste Listenposition	992 km
Arith. Mittel erste Listenposition	**1.345 km**

Tabelle 54: Experiment: Gesamtlauf mit Koordinaten

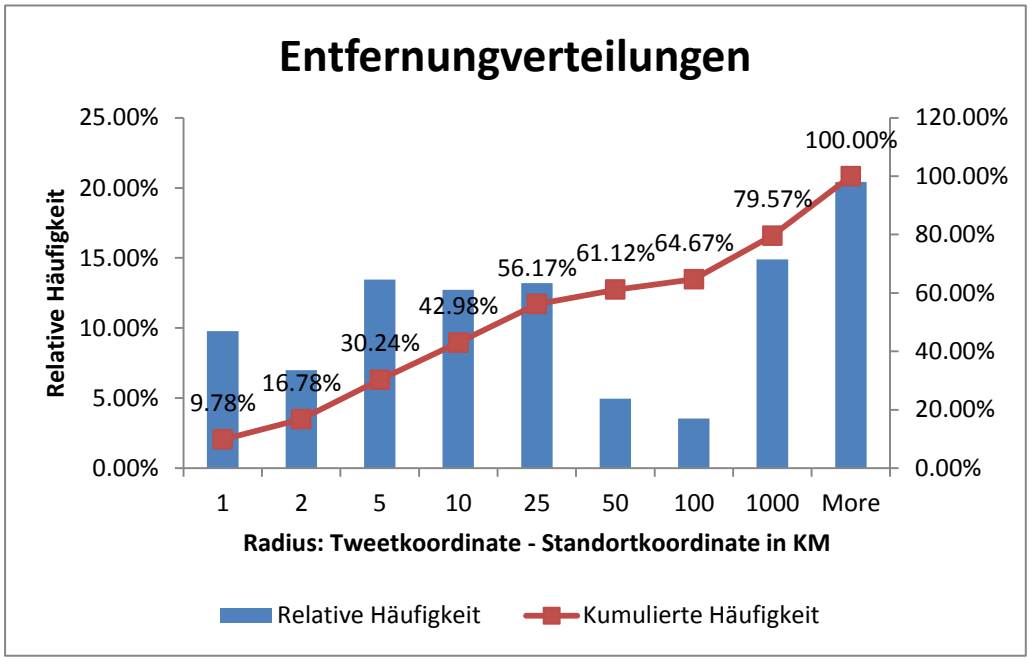


Abbildung 38: Diagramm der Entfernungsverteilungen vom Tweet zum ersten Ort der Ergebnisliste zzgl. Koordinaten

Die Tabelle 56 vergleicht die einzelnen Ergebnisse miteinander und fasst die Evaluation der Experimente zusammen.
Dabei lassen sich die Ergebnisse folgendermaßen dahingehend subsumieren, dass die Anzahl (Recall) und die Genauigkeit der Lokalisierung konträr zueinander sind. So beträgt die Spannweite des Recalls von 6,02 % bis 56,77 % und die Genauigkeit (Median) von 3,39 km bis 15,22[42] km. Abhängig vom gewählten Filter oder dessen Kombination lassen sich differenzierte Ergebnisse erzielen. Es kann jedem Filter oder dessen Kombination eine Distanzverteilung zugeordnet werden.
Eine besondere Herausforderung stellen die Ambiguität der Ortsbezeichner und somit die Wahl des richtigen Orts bei gleicher Bezeichnung dar. Erwähnenswert ist hier die Verwendung des UTC-Offsetbereichs zur Disambiguierung bei der Ortsauswahl. Mit diesem ist es realisierbar, den Einfluss der Ausreißer zu halbieren. Insbesondere ist bei der Anwendung des Ansatzes zu beachten, dass die Filter Orte herausfiltern und diesen daher generell keine Tweets zugeordnet werden können.
Mit der Kombination aller Filterungen und unter der Verwendung des UTC-Bereichs zuzüglich der Koordinaten sind 34,04 % aller Tweets mit einem Median von 9,23 km verortbar.

Experiment / Filter	**Recall insgesamt verorteter Tweets**	**Median beste Position**	**Median erste Reihe**	**Arith. Mittel beste Listen-position**	**Arith. Mittel erste Listen-position**
Baseline	50,70 %	9,24 km	18,30 km	701 km	1.329 km
Einschränkung auf Städte	48,05 %	10,15 km	16,05 km	794 km	2.834 km
Filterung von Ländern & Größer	44,94 %	7,40 km	13,62 km	667 km	1.211 km
Filterung von Ländern & Größer und Einschränkung auf Städte	42,10 %	8,11 km	12,10 km	660 km	1.249 km
Einschränkung auf Koordinaten	6,02 %	3,39 km		515 km	
Ortsauswahl unter der Verwendung des UTC-Bereichs	34,38 %	8,89 km	15,96 km	566 km	650 km
Kombination aller Filterungen unter Verwendung des UTC-Bereichs	28,27 %	7,72 km	10,73 km	520 km	581 km
Kombination aller Filterungen unter Verwendung des UTC-Bereichs zuzüglich der Koordinaten	**34,04 %**	**7,00 km**	**9,23 km**	**517 km**	**567 km**
Gesamtsystem ohne Koordinaten	50,70 %	12,04 km	17,84 km	1.036 km	1.431 km
Gesamtsystem mit Koordinaten	**56,77 %**	**10,70 km**	**15,22 km**	**992 km**	**1.345 km**

Tabelle 56: Ergebnisübersicht

[42] Ergebnis des Gesamtsystems incl. Koordinateneintragungen im Standortfeld.

6. Zusammenfassung und Ausblick

Das abschließende Kapitel fasst die Arbeit zusammen und es schließt mit einem Ausblick auf zukünftige Arbeiten.

6.1. Zusammenfassung

Soziale Medien, wie Twitter, generieren eine große Menge an Daten. Täglich werden Millionen an Tweets versendet, welche diverse verwertbare Informationen enthalten können. Um eine Nutzung der Informationen im Katastrophenfall zu ermöglichen, ist es essentiell, einen geografischen Bezug zu den Nachrichten herzustellen. Aktuell sind nur knapp ein Prozent aller Tweets mit einem Koordinatenpaar versehen. Deshalb besteht die Herausforderung darin, ein geeignetes Verfahren zu finden, Tweets geografisch zu lokalisieren. Mit der in dieser Arbeit vorgestellten Methode, Tweets auf Basis des Standortfelds zu verorten, ist es möglich, diese präziser zu lokalisieren als mit den bisher bekannten Verfahren.
Der vorgestellte Ansatz besteht darin, den geografischen Eintrag im Standortfeld als Approximation der Position des Tweets zu verwenden. Um die Präzision der Prognose der Tweetposition nachhaltig zu erhöhen, wurden weitere Verbesserungen in Form von Filtern präsentiert. Dabei ist es möglich, Länderangaben zu filtern, die Suche auf Städte einzugrenzen und ausschließlich Koordinaten aus dem Standortfeld zu betrachten sowie die einzelnen Filter zu kombinieren. Weitere Informationen aus dem Nutzerprofil, wie Zeitzone und UTC-Offset, können zur Disambiguierung bei der Ortsauswahl verwendet werden. Mit diesen ist es realisierbar, den Einfluss der Ausreißer der Prognose zu halbieren. Abhängig von der gewählten Option verbessert sich die Genauigkeit und zugleich verringert sich auch die Anzahl (Recall) der verortbaren Tweets. Das Ergebnis des Ansatzes lässt sich folgendermaßen formulieren: Die Anzahl und die Genauigkeit verhalten sich konträr zueinander. Es können 56,77 % der Tweets mit einem Median von 15,22 km geortet werden und im Vergleich dazu nur 6,02 % mit einem Median von 3,39 km. Der Anwender kann die Genauigkeit der Lokalisierung im gegebenen Rahmen somit wählen. Das entwickelte Verfahren kann somit überall dort eingesetzt werden, wo keine exakte Verortung benötigt wird und der Nutzer sich in der Nähe des eingetragenen Standorts befindet. Mögliche Einsatzszenarien sind somit bei sozialen Erdbebensensoren, bei der Meinungsforschung und bei der Verfolgung der Ausbreitung von Seuchen. Weitere Szenarien sind Katastrophen bei Veranstaltungen mit begrenztem Einzugsgebiet. Im Vergleich zu den anderen Verortungsmethoden weist das in dieser Arbeit vorgestellte Verfahren nur eine geringe Rechenkomplexität auf, ist weltweit einsetzbar und schränkt keine Tweeter in der Nachrichten- oder Beziehungsanzahl ein. Es wird nur das Standortfeld aus dem Nutzerprofil des Tweeters benötigt.

6.2. Offene Fragestellungen und Ausblick

Anknüpfpunkte für zukünftige Arbeiten ergeben sich zum einen aus den Erweiterungen des in dieser Arbeit vorgestellten Ansatzes und zum anderen aus der Kombination der in den verwandten Arbeiten vorgestellten Verfahren. Für nicht gesetzte Standortfelder lässt sich die Position über die Beziehungen[43] des Tweeters wählen und es ist zu eruieren, inwieweit diese Position mit der des Tweets übereinstimmt. Aus den Ausreißeranalysen lässt sich erschließen, dass eine Aufbereitung ungünstiger Ortsangabeformen eine Verbesserung der Verortungsgenauigkeit und des Recalls generiert. Mögliche Fehlverortungen können minimiert werden, wenn es gelingt, die vermeintlich nicht geografisch gemeinten Angaben von realen existierenden Orten zu differenzieren.
In dieser Arbeit wurde das UTC-Offset zur Disambiguierung bei der Ortsauswahl verwendet. Zu eruieren ist, inwieweit die Zeitzone sich hierfür eignen würde. Optimierungspotenziale bieten auch die Ortsangaben im Standortfeld in Kommagruppen-Form. Deren getrennte Betrachtung und Interpretation können die Suchanfrage an das geografische Lexikon verbessern.

[43] Das Verfahren zur Lokalisierung auf Basis der Nutzerbeziehungen wurde im Kapitel 3.7. vorgestellt.

Tabellenverzeichnis

Abbildungsverzeichnis

Literaturverzeichnis

1. **Maynard, Diana; Funk, Adam.** *Automatic detection of political opinions in Tweets.* Sheffield, UK, 2011.
2. **Palen, Leysia; Anderson, Kenneth M.; Mark, Gloria; Martin, James; Sicker, Douglas; Palmer, Martha and Grunwald, Dirk.** *A Vision for Technology-Mediated Support for Public Participation & Assistance in Mass Emergencies & Disasters.* Proceedings of ACM-BCS Visions of Computer Science 2010. Boulder, Irvine, 2010.
3. **Lucy Gunawan, Hani Alers, Willem-Paul Brinkman, Mark A. Neerincx.** *Distributed Collaborative Situation-Map Making for Disaster Response.* Interacting with Computers, 23(4), S. 308-316 *Delft, Netherlands,* 2011.
4. **Ushahidi**. http://ushahidi.com. Zugriff am: 01.06.2012.
5. **Twitcident**. http://twitcident.com. Zugriff am: 01.06.2012.
6. **Munro, Robert.** *Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol.* Association for Computational Linguistics. Portland, Oregon, USA, 2011.
7. **American Red Cross**. *Social Media in Disasters and Emergencies.* 2010.
8. **Handelsblatt**. http://www.handelsblatt.com/politik/deutschland/loveparade-katastrophe-kein-ende-der-ermittlungen-in-sicht/6194648.html. Deutsche Presse-Agentur GmbH, 10.02.2012. Zugriff am: 24.05.2012.
9. **Blank, Gerd**. *Loveparade im sozialen Netz - Der virtuelle Trauerzug.* http://www.stern.de/digital/online/loveparade-im-sozialen-netz-der-virtuelle-trauerzug-1587073.html. 26.07.2011. Zugriff am: 24.05 2012.
10. **Aachener Nachrichten**. http://www.aachener-nachrichten.de/lokales/euregio-detail-an/1792186. Deutsche Presse-Agentur GmbH. 24.08.2011. Zugriff am: 24.05.2012.
11. **Focus**. https://www.focus.de/digital/multimedia/krawalle-in-england-blackberry-twitter-und-facebook-im-strassenkampf_aid_653901.html. 09.08.2011. Zugriff am: 24.05.2012.
12. **Leithäuser, Johannes**. Frankfurter Allgemeine Zeitung. http://www.faz.net/aktuell/politik/ausland/krawalle-in-grossbritannien-die-fehler-der-bobbys-11112932.html. Frankfurter Allgemeine Zeitung GmbH, 08.08.2011. Zugriff am: 24.05.2012.
13. **Welt online**. http://www.welt.de/politik/ausland/article13540596/Londoner-Exzess-fordert-weiteres-Menschenleben.html. Axel Springer Verlag, 12.08.2011. Zugriff am: 24.05.2012.

14. **Lampos, Vasileios**. Flu Detector. http://geopatterns.enm.bris.ac.uk/epidemics/. University of Bristol. Zugriff am: 24.05.2012.
15. **Lampos, Vasileios; Christianini, Nello.** *Nowcasting Events from the Social Web with Statistical Learning*. ACM Transactions on Intelligent Systems and Technology, Vol.3, Article 60. Bristol, UK, September 2011.
16. **Lampos, Vasileios; Christianini, Nello.** *Tracking the flu pandemic by monitoring the Social Web.* 2nd International Workshop on Cognitive Information Processing. Bristol, UK, 2010.
17. **The Telegraph.** www.telegraph.co.uk/technology/twitter/4269765/New-York-plane-crash-Twitter-breaks-the-news-again.html. Telegraph Media Group Limited. Zugriff am: 20.06.2012.
18. **McClendon, Susannah; Robinson, Anthony C.** *Leveraging Geospatially-Oriented Social Media Communications in Disaster Response.* Vancouver, Canada, April 2012.
19. **SwiftRiver**. www.swiftly.org. Zugriff am: 06.04.2012.
20. **Tweetincident**. http://twitcident.com. Zugriff am: 04.06.2012.
21. **Abel, Fabian; Hauff, Claudia; Houben, Geert-Jan; Tao, Ke; Stronkman, Richard**. *Semantics + Filtering + Search = Twitcident Exploring Information in Social Web Streams.* HT. ACM. Milwaukee, Wisconsin, USA, 2012.
22. **Twitter**. https://twitter.com/. Twitter Inc. Zugriff am: 11.07.2012.
23. **Föll, Lars**. *Twitter 2012 - aktuelle Zahlen, Fakten und Statistiken zum Microbloggingdienst.* 26.02.2012. http://www.itrig.de/index.php?/archives/1044-Twitter-2012-aktuelle-Zahlen,-Fakten-und-Statistiken-zum-Microbloggingdienst.html. Zugriff am: 22.06.2012.
24. **Twitter Blog**. http://blog.twitter.com/2011/06/200-million-tweets-per-day.html. Twitter Inc., 30.06.2011. Zugriff am: 22.06.2012.
25. **Kinsella, Sheila; Murdock, Vanessa; O'Hare, Neil**. *"I'm Eating a Sandwich in Glasgow": Modeling Locations with Tweets.* SMUC. ACM. Glasgow, Scotland, UK, 2011.
26. **Hale, Scott A., Gaffney, Devin and Graham, Mark.** *Where in the world are you? Geolocation and language identification in Twitter.* Oxford, United Kingdom,
27. **Cheng, Zhiyuan; Caverlee, James; Lee, Kyumin.** *You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users.* CIKM. ACM. Toronto, Ontario, Canada, 2010.
28. **Takahashi, Tetsuro; Abe, Shuya; Igata, Nobuyuki.** *Can Twitter Be an Alternative of Real-World Sensors?*. Springer-Verlag Berlin Heidelberg, 2011.
29. **Watanabe, Kazufumi; Ochi, Masanao; Okabe, Makoto; Onai, Rikio.** *Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs.* CIKM. ACM. Glasgow, Scotland, UK, 2011.
30. **Boyd, Danah; Scott, Golder; Lotan, Gilad.** *Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.* HICSS-43. IEEE. Kanuai, HI, 2010.
31. **Einat Amitay, Nadav Har´El, Ron Sivan, Aya Soffer.** *Web-a-Where: Geotagging Web Content.* SIGIR. ACM. Sheffield, South Yorkshire, UK, 2004.
32. **Lieberman, Michael D.; Samet, Hanan; Sankaranayananan, Jagan.** *Geotagging: Using Proximity, Silbing, and Prominence Clues to Understand Comma Groups.* GIR. ACM. Zurich, Switzerland, 2010.
33. **Smith, David A.; Crane, Gregory.** *Disambiguating Geographic Names in a Historical Digital Library.* Proc. of the 5th European Conf. on Research and Advanced Technology for Digital Libaries. Medford, MA, USA, 2001
34. **Woodruff, Allison Gyle; Plaunt, Christian**. *GIPSY: Automated Geographic Indexing of Text Documents*. Journal of the American Society for Information Science, 45(9) S. 645-655, 1994.
35. **Wikipedia.** - *Artikel Russland.* https://de.wikipedia.org/wiki/Russland#. Zugriff am: 06.08.2012.
36. **MetaCarta**. http://www.metacarta.com. MetaCarta, a Division of Qbase. Zugriff am: 20.06.2012.
37. **The Google Geocoding API**. https://developers.google.com/maps/documentation/geocoding/.Google Inc.. Zugriff am: 20.06.2012.
38. **Yahoo! PlaceFinder**. http://developer.yahoo.com/geo/placefinder. Yahoo Inc.. Zugriff am: 12.06.2012.

39. **Gisgraphy** - http://www.gisgraphy.com. Zugriff am: 15.06.2012.
40. **USGS Geographic Names Information System (GNIS).** http://geonames.usgs.gov. Zugriff am: 05.06.2012.
41. **United Nations department of economic and social.** http://unstats.un.org/unsd. Zugriff am: 05.06.2012.
42. **OpenStreetMap: Die freie Wiki-Weltkarte.** ww.openstreetmap.de. FOSSGIS e.V. Zugriff am: 12.06.2012.
43. **World Gazetteer.** http://world-gazetteer.com. Zugriff am: 06 05, 2012.
44. **Getty Thesaurus of Geographic Names.** www.getty.edu/research/tools/ - Vocabularies/tgn. Getty Research Insitute. Zugriff am: 20.06.2012.
45. **ISO - International Organisation for Standardization**. www.iso.org/iso/country_codes/iso_3166_code_lists.htm. Zugriff am: 08.06.2012.
46. **Wikipedia - Die freie Enzyklopädie**. https://de.wikipedia.org. Zugriff am: 20.06.2012.
47. **Webservice von Geonames.** http://api.geonames.org/.Zugriff am: 04.07. 2012.
48. **GeoNames - Search Webservice.** www.geonames.org/export/geonames-search.html. Zugriff am: 10.07.2012.
49. **Larson, Martha; Soleymani, Mohammad, Serdykov, Pavel.** *Automatic Tagging and Geotagging in Video Collections and Communities.* ICMR. ACM. Trento, Italy, 2011
50. **Marsh, Elaine; Perzanowski, Dennis.** *MUC-7 EVALUATION OF IE TECHNOLOGY: Overview of Results.* 1998.
51. **Message Understanding Conference**. Proceedings MUC-7 Table of Contents. www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html#named. Zugriff am: 21.06.2012.
52. **Charniak, Eugene.** *Statistical Techniques for Natural Language Parsing*. Brown University, August 1997.
53. **Gelernter, Judith; Mushegian, Nikolai.** *Geo-parsing Messages from Microtext.* Transaction in GIS, 25(6): S. 753-773. Blackwell Publishing Ltd., 2011.
54. **OpenCalais**. http://www.opencalais.com/. Thomson Reuters. Zugriff am: 12.06.2012.
55. **Woodruff, Allison Gyle; Plaunt, Christian.** *GIPSY: Automated Geographic Indexing of Text Documents*. Journal of the American Society for Information Science, 45(9) S. 645-655, 1994.
56. **Hecht, Brent; Gergle, Darren.** *On the "Localness" of User-Generated Content.* CSCW. ACM. Savannah, Georgia, USA, 2010.
57. **Hecht, Brent; Hong, Lichan; Suh ,Bongwon; Chi, Ed h..** *Tweets from Justin Bieber´s Heart: The Dynamics of the "Location" Field in User Profiles.* CHI. ACM. Vancouver, BC, Canada, 2011.
58. **Paradesi, Sharon.** *Geotagging Tweets Using Their Content.* Proccedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference. Florida, USA, 2011.
59. **Yohei Ikawa, Miki Enoki, Michiaki Tatsubori.** *Location Inference using Microblog Messages. WWW2010. ACM.* Lyon, France, 2012.
60. **Eisenstein, Jacob; O´Connor, Brendan; Smith, Noah A.; Xing, Eric P..** *A Latent Variable Model for Geographic Lexical Variation.* Pittsburgh, PA 15213, USA
61. **Blei, David M.; Ng, Andrew Y.; Jordan, Michael.** *Latent Dirichlet Allocation*. Journal of Machine Learning Research 3, S. 993-1022. 2003.
62. **Davis Jr., Clodoveu A. ; Papa, Gisele L.; Oliveira, Diogo Renno Rocha de; Arcanjo, Filipe de L..** *Inferring the Location of Twitter Messages Based on User Relationships.* Transactions in GIS, 25(6): S. 735-751. Blackwell Publishing Ltd., 2011.
63. **Hurst, Matthew; Siegler, Matthew; Glance, Natalie.** *On Estimating The Geographic Distibution of Social Media.* ICWSM. Boulder, Colorado, USA, 2007.

64. **Backstrom, Lars; Sun, Eric; Marlow, Cameron.** *Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. WWW 2010.* ACM. Raleigh, North Carolina, USA, 2010.
65. **Kwak, Haewoon; Lee, Changhyun; Park, Hosung; Moon, Sue.** *What is Twitter, a Social Network or a News Media? WWW 2010.* ACM. Raleigh, North Carolina, USA, 2010.
66. **McGee, Jeffrey; Caverlee, James; Cheng, Zhiyuan.** *A Geographic Study of Tie Strength in Social Media.* CIKM. ACM. Glasgow, Scotland, UK, 2011.
67. **Salcatore Scellato, Anastasios Noulas.** *Socio-spatial Properties of Online Loction-based Social Networks.* Association for the Advancement of Artificial Intelligence, 2011.
68. **Krishnamurthy, Balachander; Gill, Phillipa; Arlitt, Martin.** *A Few Chirps About Twitter.* WOSN. ACM. Seattle, Washington, USA, 2008.
69. Flicker. www.flicker.com. Yahoo Inc.. Zugriff am: 03.06.2012.
70. **McCurley, Kevin S..** *Geospatial Mapping and Navigation of the Web. WWW 2010.* ACM. Hong Kong, 2001.
71. **Tweak the Tweet**. *Project EPIC - Empowering the Public with Information in Crisis*. http://epic.cs.colorado.edu/. Zugriff am: 04.6.2012.
72. **Java, Akshay; Song, Xiaodan**. *Why We Twitter: Understanding Microblogging Usage and Communities.* WEBKDD & 1st SNA-KDD Workshop August 2010. ACM. San Jose California, USA, 2007.
73. **Hecht Brent, Emily Moxley.** *Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge*. COSIT 09, LNCS 5756, S. 88-105. Springer-Verlag Berlin Heidelberg, 2009.
74. **Duden**. www.duden.de/rechtschreibung/Homonym. Zugriff am: 13.06.2012.
75. **Twitpic**. http://twitpic.com/135xa. Zugriff am: 20.06.2012.
76. **Wirtschafts Woche**. http://blog.wiwo.de/ungedruckt/2012/04/06/infografik-der-woche-deutschland-ist-twitter-entwicklungsland-noch/. Handelsblatt GmbH. Zugriff am: 22.06.2012.
77. **Ogg, Erica**. http://gigaom.com/2011/09/08/twitter-ceo-we-have-100m-active-users/. 08.09.2011. Zugriff am: 22.06.2012.
78. **Terpstra, Teun; Stronkman, R.; de Vries, A.; Paradies, G.L.**. *Towards a realtime Twitter analysis during crises for operational crisis management*. Proceedings of the 9th International ISCRAM Conference. Vancouver, Canada, April 2012.
79. **KNAUERS LEXIKON A-Z**. Droemersche Verlagsanstalt Th. Knauer Nachf., München, 1991.
80. **Lohmann, Steffen; Burch, Michael; Schmauder, Hansjörg; Weiskopf, Daniel**. *Visual Analysis of Microblog Content Using Time-Varying Co-occurrence Highlighting in Tag Clouds.* Capri Island, Italy, 2012.
81. **Kamianets, Wolodymyr**. *Zur Einteilung der deutschen Eigennamen. Grazer Linguistische Studien.* Graz, 2000.
82. Brockhaus Enzyklopädie. F.A. Brockhaus. Leipzig - Mannheim, 2006.
83. **Witten, Ian H.; Frank, Eibe**. *Data Mining Practical Machine Learning Tools and Techniques.* Morgan Kaumann Publishers, 2005.