

Universität Leipzig
Text Mining and Retrieval
Studiengang Informatik

Korpuskonstruktion und Evaluierungsschema zur Erkennung von Extremereignissen über Twitter

Bachelorarbeit

Hannes Winkler

1. Gutachter: Jun.-Prof. Dr. Martin Potthast
2. Gutachter: Dr. Andreas Niekler

Mentor: Matti Wiegmann

Datum der Abgabe: 27. März 2020

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Leipzig, 27. März 2020

.....
Hannes Winkler

Zusammenfassung

Ereigniserkennung über soziale Medien ermöglicht das Erstellen von Warnsystemen zum Registrieren und Nachverfolgen natürlicher und anthropogener Katastrophen. Insbesondere der Mikroblogging-Dienst Twitter ist für die Erstellung eines solchen Systems geeignet, da neue Nachrichten über kein anderes Netzwerk schneller publiziert und geteilt werden. State-of-the-Art-Ansätze zur Ereigniserkennung über Twitter unterscheiden sich in der Anwendbarkeit zur Erkennung von Krisen verschiedener Typen (Übertragbarkeit), der Verarbeitung von Tweets und den Datensätzen, die zur Evaluation genutzt wurden. Aufgrund dieser Heterogenität werden in dieser Bachelorarbeit die vier Ansätze Keywords, Classification, Clustering und Burstiness nachgebildet und hinsichtlich ihrer Übertragbarkeit und Leistungsfähigkeit einheitlich evaluiert. Hierfür wurde erstmalig durch die Recherche in Krisendatenbanken und die manuelle Annotation von Tweets ein Textkorpus konstruiert, der als neuartiger Standard zur Evaluation von Erkennungsalgorithmen angesehen werden kann. Dafür wurde ein Datensatz mit 292.981.506 Tweets genutzt, der von der Bauhaus-Universität Weimar bereitgestellt wurde und sich aus 165 Zeiträumen von 2011 bis 2019 zusammensetzt. Für die manuelle Annotation erkenntnisrelevanter Tweets wurden zunächst zehn Online-Datenbanken nach Krisen durchsucht, danach die drei Desastertypen Erdbeben, Flut und Amoklauf ausgewählt und anschließend die ersten Tweets von Betroffenen markiert. So konnten sechs Tweets über Erdbeben, drei Tweets über Fluten und vier Tweets über Amokläufe ermittelt werden. Mit Hilfe dieser 13 Desaster wurden die Erkennungsalgorithmen evaluiert, indem die Übertragbarkeit geprüft und die durchschnittlichen F1-Werte berechnet wurden. Die Auswertung ergab, dass sowohl der Keywords- (F1-Wert 0,3132) und der Classification- (F1-Wert 0,3050) als auch der Burstiness-Algorithmus (F1-Wert 0,2873) übertragbar sind. Lediglich der Clustering-Algorithmus (F1-Wert 0,4364) konnte keine Flut erkennen. Die Analyse der korrekt sowie fälschlich erkannten Ereignisse zeigte, dass die drei primären Ursachen für fälschlich erkannte Ereignisse auf Spam (45 %), Tweets von Bots (25 %) und Nachrichtenmeldungen (17 %) zurückzuführen sind. Es zeigte sich, dass der Classification-Algorithmus das größte Potenzial für die weitere Forschung birgt, da dieser den Kontext eines Tweets zur Relevanzprüfung einbezieht. Zukünftige Aufgaben umfassen die Reduzierung fälschlich erkannter Ereignisse, die Geolokalisierung betroffener Twitter-Nutzer und die Ereigniserkennung in Echtzeit.

Inhaltsverzeichnis

1	Einleitung	1
2	Related Work	3
3	Korpuskonstruktion	5
3.1	Grundlagen der Datenakquise	5
3.2	Korpusaufbau	6
3.3	Datenbanksuche	7
3.4	Annotation erkenntnisrelevanter Tweets	9
3.4.1	Prämissen und Durchführung	9
3.4.2	Beobachtungen	11
4	Methodik	13
4.1	Grundlagen der Ereigniserkennung	13
4.2	Algorithmen zur Ereigniserkennung	14
4.2.1	Keywords	14
4.2.2	Classification	15
4.2.3	Clustering	16
4.2.4	Burstiness	18
4.3	Experimente	19
5	Ergebnisse	21
6	Diskussion	28
7	Zusammenfassung	30
	Literaturverzeichnis	31
A	Wortlisten	38
B	Auswertung der Erkennungsalgorithmen	39
C	Fälschlich erkannte Ereignisse	42

Kapitel 1

Einleitung

Das Risiko materieller und menschlicher Verluste durch natürliche und anthropogene Katastrophen kann mit Hilfe von Frühwarnsystemen verringert werden. Doch obwohl Messsysteme kontinuierlich weiterentwickelt und verbessert werden, können sie aus Kostengründen nicht flächendeckend eingesetzt werden. Außerdem umfasst das Spektrum der Sensorik keine Desaster wie Epidemien und terroristische Anschläge. Um an Informationen aus solchen Krisen zu gelangen, sind Meldungen von involvierten Personen nötig. Hierbei bergen im Zeitalter der digitalen Vernetzung soziale Medien großes Potenzial.

Seit seiner Gründung 2006 hat sich der Mikroblogging-Dienst Twitter zu einer der wichtigsten Kommunikationsplattformen für Nachrichten und Trends etabliert. Mittlerweile zählt das soziale Netzwerk 145 Millionen aktive Nutzer¹ und 500 Millionen verfasste Kurznachrichten (Tweets) täglich (Twitter Inc. [2019], Twitter Inc. Business [2019]). Twitter lässt sich vor allem durch Prägnanz und Aktualität charakterisieren. In nur 280 verfügbaren Zeichen (seit November 2017, davor 140) kommentieren Nutzer das Zeitgeschehen und teilen persönliche Erlebnisse (Java et al. [2009]). Beachtlich ist dabei, dass Berichte über neue Ereignisse oftmals vor Meldungen anderer Medien im Twitterstream erscheinen (Newman [2011], Hu et al. [2012]). Damit bilden Tweets einen vielversprechenden Ausgangspunkt für die Forschung zur Ereigniserkennung.

Wissenschaftler haben dieses Potenzial erkannt und auf die Erkennung sowie das Tracking von natürlichen und anthropogenen Katastrophen über soziale Netzwerke, insbesondere Twitter, übertragen (Gruber et al. [2015]). Anfänglich wurden vor allem Erdbeben betrachtet, da Informationen nicht in Echtzeit an Betroffene versendet werden konnten. Doch heutzutage existieren Sensorennetzwerke, die bereits kleinste Erderschütterungen registrieren und

¹Aus Gründen der besseren Lesbarkeit wird in dieser Arbeit für Personenbezeichnungen das generische Maskulin verwendet. Dieses schließt jederzeit und ausnahmslos alle Geschlechter gleichermaßen ein.

automatisierte Nachrichten über Twitter veröffentlichen. Daher ist es mittlerweile notwendig, zum einen Botnachrichten zu verifizieren und zum anderen die ermittelbaren Ereignistypen der Erkennungsalgorithmen zu erweitern, sodass andere Desaster ebenfalls erkannt werden können. Des Weiteren sind State-of-the-Art-Ansätze für Echtzeitauswertungen ungeeignet, da sie feste Zeitintervalle betrachten und bei simultan auftretenden Ereignissen keine differenzierten Informationen zum zeitlichen Verlauf und Raum extrahieren können. Aufgrund dieser Heterogenität ist es notwendig, einen einheitlichen Überblick über die Leistungsfähigkeit und insbesondere Übertragbarkeit bereits existierender Algorithmen zu erstellen.

Dazu soll in dieser Bachelorarbeit zunächst ein Korpus, bestehend aus 292.981.506 Tweets, in Bezug auf die drei Ereignistypen Erdbeben, Flut und Amoklauf manuell annotiert werden, sodass anschließend die vier Erkennungsalgorithmen Keywords nach Olteanu et al. [2014], Classification nach Sakaki et al. [2010], Clustering nach Kaleel and Abhari [2014] und Burstiness nach Kleinberg [2002] hinsichtlich ihrer Leistungsfähigkeit evaluiert werden können. Hierbei soll der annotierte Korpus als Standard dienen, welcher neben der Evaluation auch Einblicke in die Herausforderungen der Ereigniserkennung über Twitter erlaubt. Auf Grundlage dieser Arbeit soll es möglich sein, Aussagen über die Übertragbarkeit der Algorithmen auf bisher kaum beachtete Ereignistypen zu formulieren, einen Algorithmus für die weitere Forschung zu empfehlen und Verbesserungsvorschläge zu nennen.

Zum Erreichen dieser Ziele wurden zunächst die im Korpus vorhandenen Zeiträume aufgelistet. Daraufhin wurden zehn Online-Datenbanken nach Extremereignissen in diesen Intervallen durchsucht, wobei Einträge desselben Typs, die zeitlich und raumbezogen eng beieinanderliegen, zu einem Ereignis aggregiert wurden. Anschließend wurde der Korpus mit Hilfe von katastrophentypischen Keywords gefiltert, mit den Einträgen aus den Datenbanken manuell abgeglichen und die ersten relevanten Tweets markiert. So ließen die vom Korpus abgedeckten Desaster ermittelt. Daraufhin wurden die vier oben genannten Algorithmen auf den gesamten Korpus angewandt und die Ergebnismenge mit dem Resultat des vorherigen Schritts abgeglichen.

Die Evaluation zeigte, dass der Keywords- (F1-Wert 0,3132), der Classification- (F1-Wert 0,3050) und der Burstiness-Algorithmus (F1-Wert 0,2873) übertragbar sind. Der Clustering-Algorithmus (0,4364) erzielte zwar den besten durchschnittlichen F1-Wert, scheiterte jedoch bei der Erkennung von Fluten. Die besten Erkennungszeiten erreichten der Keywords- (116 Sekunden) und der Classification-Algorithmus (119 Sekunden), während die schlechtesten Zeiten auf den Burstiness-Algorithmus (772 Sekunden) und den Clustering-Algorithmus (3329 Sekunden, ohne Fluten) entfielen. Es zeigte sich, dass der Classification-Algorithmus für die weitere Forschung das größte Potenzial birgt.

Kapitel 2

Related Work

Das Erkennen sowie Nachverfolgen von Ereignissen über das Internet wurde erstmals von Allan et al. [1998] als neue Disziplin des Information Retrieval beschrieben. Die ersten teilweise unüberwachten und retrospektiven Ansätze zur Ereigniserkennung (Yang et al. [1999], Yang et al. [2000]) beschränkten sich dabei auf Clustering- und Klassifikationsalgorithmen. Um einen besseren Überblick über den zeitlichen Verlauf von Ereignissen zu erlangen, untersuchte Kleinberg [2002] ungewöhnlich häufige Vorkommnisse von Begriffen in Streams, sogenannte Bursts. Darauf bauten Fung et al. [2005] auf und entwickelten einen der ersten Ansätze zur Ereigniserkennung, der ohne manuell annotierte Trainingsdaten auskommt.

Mit steigender Popularität sozialer Medien erweiterten sich die Einsatzfelder und Möglichkeiten der Ereigniserkennung. Hierbei gelten die Forschungsarbeiten von Rattenbury et al. [2007], Zhao and Mitra [2007] und Sayyadi et al. [2009] als richtungweisend, denn sie analysierten erstmals Daten aus Blogs und Internetforen wie Flickr. Kolari et al. [2006] und Weng and Lee [2011] beschreiben die damit einhergehenden Herausforderungen, wie beispielsweise die enorme Menge an irrelevanten Postings oder Spam. Sankaranarayanan et al. [2009], Becker et al. [2011] sowie Osborne et al. [2014] untersuchen und bestätigen die Beschaffenheit von Twitter zum Teilen von Neuigkeiten und zum Erkennen von Ereignissen in Echtzeit. Daraus ergibt sich die Chance, auf Basis von Tweets und in Kombination mit Twitter-spezifischen Features wie Hashtags (Feng et al. [2015]) sowie Metadaten (Cai et al. [2015]) Frühwarn- und Trackingsysteme für Katastrophen zu entwickeln, die neben dem Ereignistyp auch den Raum einbeziehen (De Longueville et al. [2009]).

Sakaki et al. [2010] fokussieren sich in ihrer Forschung auf das Erkennen von Erdbeben, wobei Twitter-Nutzer als „soziale Sensoren“ angesehen werden. Auch Poblete et al. [2018] betrachten Erdbeben und bauen dabei auf dem Algorithmus von Earle et al. [2012] auf, welcher Bursts des Wortes „earthquake“

observiert. Das von Li et al. [2012] erstellte Twitter-based Event Detection and Analysis System kann neben Erdbeben auch Verbrechen erkennen. Ebenfalls umfassender anwendbar ist das von McCreadie et al. [2016] vorgestellte Emergency Analysis Identification and Management System. Als Grundlage wählten die Autoren den von Broder [1997] beschriebenen Ansatz des Locality Sensitive Hashing und unterstützen aktuell das Erkennen von Fluten sowie öffentlichen Demonstrationen. Um Disaster nahezu universell registrieren zu können, beschreiben Olteanu et al. [2014] und Temnikova et al. [2015] die Erstellung einer Wort- und Wortgruppenlisten mit katastrophentypischen Begriffen zum Filtern von Tweets. Da die meisten Algorithmen nur Text und Metadaten berücksichtigen, inkludieren Alqhtani et al. [2015] sowie Takamu et al. [2015] zur Erhöhung der Genauigkeit zusätzlich Bilder.

Zur Systematisierung dieser Ansätze führten Farzindar and Khreich [2013] und Cordeiro and Gama [2016] Metastudien durch. Dabei analysierten die Autoren, ob Algorithmen überwacht oder unüberwacht laufen und retrospektiv oder in Echtzeit eingesetzt werden können. Darüber hinaus vergleichen Weiler et al. [2016] die Rechenleistungen und Sato et al. [2018] die Zuverlässigkeit der Algorithmen. Saeed et al. [2019] beschreiben die von den Algorithmen einbezogenen Twitter-Features und die von den Forschern verwendeten Testdatensätze.

Kapitel 3

Korpuskonstruktion

In diesem Kapitel wird der genutzte Textkorpus, die Analyse von Krisendatenbanken zum Ermitteln abgedeckter Extremereignisse und die Annotation erkenntnisrelevanter Tweets beschrieben. Das Ergebnis sind 13 markierte Tweets, die als Ausgangspunkt für die Evaluation dienen. Beobachtete Auffälligkeiten während der Analyse und Annotation wurden dokumentiert.

3.1 Grundlagen der Datenakquise

Auf Twitter können sich Nutzer untereinander folgen, um Tweets zu teilen und zu lesen. Dabei können Tweets in drei Kategorien eingeteilt werden. Ein Tweet wird als Statusmeldung bezeichnet, wenn kein anderer Tweet oder Nutzer referenziert wird, als Retweet, wenn eine Statusmeldung geteilt wird, und als Antwort, wenn eine Statusmeldung oder ein Retweet kommentiert wird. Nutzer können einstellen, ob Tweets öffentlich oder nur für Follower einsehbar sind.

Über die Twitter-Programmierschnittstelle (API) können öffentliche Tweets als sogenannte Tweet-Objekte (Twitter Inc. [2020]) sowohl rückwirkend als auch in Echtzeit im JavaScript Object Notation Datenformat abgerufen werden. Die Objekte beinhalten neben dem vom Autor verfassten Text bis zu 35 Attribute. So erhält jeder Tweet eine einzigartige ID, einen Zeitstempel im UTC-Format, der auf die Millisekunde genau ist, und die einzigartige ID des Autors. Ein weiteres Attribut beinhaltet, wenn vom Nutzer angegeben, geographische Informationen wie beispielsweise Koordinaten und Ortsbezeichnungen. Hierbei ist anzumerken, dass das Geotag nicht zwangsläufig Rückschlüsse über den exakten Aufenthaltsort des Autors erlaubt, da bereits die Angabe des Landes genügt. Wenn ein Tweet retweeted wird, werden zwei Tweet-Objekte verknüpft. So beinhaltet jedes Retweet-Objekt das vollständige referenzierte Tweet-Objekt. Bei Antworten wird die ID der ursprünglichen Statusmeldung im Tweet-Objekt hinterlegt.

Für die Datenakquise zur Ereigniserkennung ist es essentiell, Retweets und Antworten auszuschließen, denn sie enthalten keine neuen Informationen und können zur fehlerhaften Erkennung von Ereignissen führen, wenn Algorithmen die Anzahl von Tweets in einem Intervall auswerten. Außerdem ist es nicht nötig, alle Metadaten zu übernehmen, da Attribute wie beispielsweise die Angabe, ob ein Tweet sensible Inhalte beinhaltet, für die Ereigniserkennung nebensächlich sind.

3.2 Korpusaufbau

Für diese Arbeit wird ein Textkorpus genutzt, den die Mitarbeiter der Bauhaus-Universität Weimar erstellten. Dieser beinhaltet 292.981.506 Tweet-Objekte (zur besseren Lesbarkeit im weiteren Verlauf als Tweets bezeichnet), welche die Attribute Text, Tweet-ID, Autor-ID und Zeitstempel beinhalten. Die Tweets erstrecken sich vom 5. Januar 2011 bis zum 15. Oktober 2019 und decken 165 Zeiträume zwischen 20 Minuten und 24 Stunden ab. Als Vorlage für 153 dieser Zeiträume diente die von CNN Labs [2018] erstellte Datenbank über Amokläufe an Bildungseinrichtungen. Dabei wurden die Zeitpunkte aus der Datenbank in 20-minütige Intervalle umgewandelt, indem fünf Minuten davor und 15 Minuten danach hinzugefügt wurden.

Tabelle 3.1: Desaster im Korpus, exklusive Amokläufe an Bildungseinrichtungen.

Ereignis	Ort	Beginn (UTC)	Ende (UTC)
Eisenbahnunfall	Port Richmond	13.05.2015 00:53	13.05.2015 03:23
Amoklauf	Orlando	12.06.2016 05:52	12.06.2016 09:24
Erdbeben	Kaikoura	13.11.2016 10:32	13.11.2016 16:32
Amoklauf	Las Vegas	02.10.2017 04:55	02.10.2017 05:45
Schlammlawine	Santa Babara	09.01.2018 11:30	09.01.2018 13:30
Hurrikan	Ostküste USA	14.09.2018 11:00	15.09.2018 11:00
Gasexplosion	Noble County	18.01.2019 15:10	18.01.2019 18:10
Tornado	Südosten USA	03.03.2019 18:25	04.03.2019 01:55
Erdbeben	Südkalifornien	06.07.2019 03:00	06.07.2019 05:00
Ausschreitung	Hongkong	13.08.2019 13:45	13.08.2019 16:45
Erdbeben	Texas	10.09.2019 03:55	10.09.2019 05:55
Erdbeben	Kalifornien	15.10.2019 19:12	15.10.2019 21:12

Die übrigen zwölf Zeiträume sind in Tabelle 3.1 gelistet und beinhalten den Eisenbahnunfall von Port Richmond, Philadelphia, am 12. Mai 2015, die Erdbeben in Kaikoura, Neuseeland, am 14. November 2016, in Südkalifornien am

6. Juli 2019, in Texas am 10. September 2019 und in Pleasant Hill, Kalifornien, am 15. Oktober 2019, die Amokläufe in Orlando am 12. Juni 2016 und in Las Vegas am 2. Oktober 2017, die Schlammlawine nahe Santa Barbara, Kalifornien, am 9. Januar 2018, die Ankunft des Hurrikans „Florence“ an der Ostküste der Vereinigten Staaten am 14. September 2019, die Explosion einer Gasleitung in Noble County, Ohio, am 18. Januar 2019, die Tornado-Welle im Südosten der USA am 3. März 2019 sowie die Ausschreitungen am Flughafen von Hongkong am 13. August 2018.

Alle Tweets wurden rückwirkend über die Twitter API abgefragt, wobei Retweets und Antworten herausgefiltert wurden. Der zusammengestellte Korpus deckt eine Teilmenge des öffentlichen Twitterstreams innerhalb der genannten Zeiträume ab. Deswegen kann davon ausgegangen werden, dass im Korpus auch Tweets über andere Desaster enthalten sind.

3.3 Datenbanksuche

Zur Ermittlung weiterer Desaster wurden Online-Datenbanken von fünf staatlichen Institutionen, einer Universität und einer Non-Profit-Organisation analysiert und nach den 165 Zeiträumen aus dem vorgegebenen Korpus durchsucht. Das Resultat ist in Tabelle 3.2 dargestellt. Es konnten insgesamt 3.007 Einträge extrahiert werden.

Tabelle 3.2: Ergebnis der Suche nach Krisen, die innerhalb der Korpuszeiträume aktiv waren.

Desaster	Datenbank-anbieter	Gebiet	Treffer	Genauigkeit
Universell	ADRC	Global	77	Tag
Lawine	ESSL	Europa	8	Minute/Stunde
Erdbeben	USGS	Global	2183	Zehntelsekunde
Flut	DFO	Global	333	Tag
Vulkan	NOAA	Global	6	Tag
Tornado	NOAA	USA	18	Minute
Sturm	NOAA	USA	86	Minute
Hagel	NOAA	USA	67	Minute
Erdrutsch	NASA	Global	107	Tag
Amoklauf	GVA	USA	122	Tag

Die globale und manuell erstellte „GLIDENumber“ Datenbank des Asian Disaster Reduction Centre (ADRC [2020]) beinhaltet taggenaue Einträge zu natürlichen und anthropogenen Krisen, die mit Beschreibungen zur Ursache und zum Verlauf versehen sind. Es konnten 77 Ereignisse ermittelt werden, wobei sich die drei häufigsten Kategorien aus Fluten (23), Epidemien (12) und Erdbeben (6) zusammensetzten. Jedoch stellte sich bei der Analyse der gefundenen Ereignisse heraus, dass raumzeitliche Angaben in der „GLIDENumber“ Datenbank approximiert sind und komplett auf Einträgen aus anderen Datenbanken beruhen, wodurch möglicherweise die Detailliertheit der Angaben gemindert ist und die gelisteten Extremereignisse unvollständig sind.

Die europäische Unwetterdatenbank „European Severe Weather Database“ der gemeinnützigen Forschungsorganisation European Severe Storms Laboratory (ESSL [2020]) beinhaltet unter anderem Einträge zu Starkregen, Tornados, Hagel- und Schneestürmen sowie Lawinen. Jeder Eintrag basiert auf Nachrichten- und Wetterdienstmeldungen und ist sowohl mit einer stundengenauen Zeitangabe als auch mit Koordinaten versehen. Allerdings wird für die Datenbank keine öffentliche Programmierschnittstelle zur Verfügung gestellt und manuelle Abfragen über die Webseite sind auf 25 Tabelleneinträge pro Zeitraum limitiert. Deswegen wurde die Suche auf Lawinen beschränkt. Es konnten acht Einträge manuell ermittelt werden.

Der „Earthquake Catalog“ der US-amerikanischen Behörde United States Geological Survey (USGS [2020]) listet Erdbeben weltweit, bietet Suchparameter an und gibt Resultate in Form einer csv-Datei zurück. Da die Datenbank auf Sensorendaten basiert, befinden sich in ihr auch Einträge über Mikrobeben mit Magnituden zwischen 0 und 4. Jeder Eintrag ist mit einem sekundengenauen Zeitstempel, den Koordinaten des Epizentrums und der Magnitude des Erdbebens versehen. Eine uneingeschränkte Abfrage lieferte 2.183 Treffer.

Das „Global Active Archive of Large Flood Events“ der universitären Forschungsgruppe Dartmouth Flood Observatory (DFO [2020]) beinhaltet Fluten weltweit und reicht bis ins Jahr 1985 zurück. Die Einträge basieren auf Berichten der FloodList [2020] und auf Nachrichtenmeldungen. Da Fluten im Gegensatz zu Erdbeben langsam entstehen und sich über einen längeren Zeitraum erstrecken, sind in der Datenbank sowohl Startzeitpunkte als auch Endzeitpunkte taggenau gespeichert. Hierbei ist anzumerken, dass diese Zeitpunkte, insbesondere die Endzeitpunkte, bei der Datenaufnahme approximiert werden, wenn die Administratoren keine Zeitangaben in den Berichten vorfinden. Die Suchanfrage lieferte eine xlsx-Datei mit 333 Treffern, wobei jede Flut durch einen taggenauen Zeitraum, Koordinaten, eine Ortsbezeichnung und die Ursache beschrieben ist.

Die globale „The Significant Volcanic Eruption Database“ des National Centers for Environmental Information der US-amerikanischen Behörde Natio-

nal Oceanic and Atmospheric Administration (NOAA [2020]) listet über 800 Vulkanausbrüche. Die Suchanfrage lieferte eine tsv-Datei mit sechs Treffern, wobei jeder Eintrag einen taggenauen Zeitstempel, Koordinaten und Ortsbezeichnungen enthält. Es liegen keine Angaben zur Datenerhebung vor.

Die Suchanfragen nach Tornados, Stürmen und Hagelunwettern in der nationalen Datenbank des Storm Prediction Center der National Oceanic and Atmospheric Administration (NOAA [2019]) lieferten 18, 86 und 67 Treffer. Alle Einträge beruhen auf Sensorendaten des nationalen Wetterdienstes National Weather Service der USA (NWS [2020]), beinhalten Koordinaten sowie minutengenaue Zeitstempel und werden als csv-Datei ausgegeben. Allerdings ist die Datenbank nicht aktuell, denn der letzte Eintrag ist vom 31. Dezember 2018.

Der „Global Landslide Catalog“ der National Aeronautics and Space Administration (NASA [2015]) beinhaltet weltweite Einträge zu Schlammlawinen, Erdbeben und Steinschlägen. Die Einträge beruhen auf Meldungen aus Nachrichtenportalen sowie wissenschaftlichen Fachzeitschriften und setzen sich aus Koordinaten, Ortsbezeichnungen, taggenauen Zeitstempeln und den Ursachen zusammen. Jedoch ist die Datenbank nicht aktuell, denn der letzte Eintrag ist von März 2015. Die Suchanfrage lieferte eine csv-Datei mit 107 Treffern.

Die Non-Profit-Organisation Gun Violence Archive (GVA [2020]) dokumentiert Waffengewalt in den USA seit 2014. In der gleichnamigen Datenbank sind aktuell 292.042 Einträge über gewalttätige Polizeieinsätze, Einbrüche und Amokläufe gelistet, die auf Daten aus Polizei- und Regierungsberichten sowie Nachrichtenmeldungen beruhen. Da keine Suchparameter zur Verfügung gestellt werden und Anfragen auf 500 Treffer limitiert sind, wurden nur Einträge über Amokläufe mit mindestens vier Opfern, das heißt verletzte und getötete Personen, angefragt. Es wurde eine csv-Datei mit 122 Treffern zurückgegeben, wobei sich die Einträge aus einem taggenauen Zeitstempel, der Bezeichnung des Ortes inklusive Straßename und Informationen zur Anzahl der Verletzten sowie Toten zusammensetzen.

3.4 Annotation erkenntungsrelevanter Tweets

3.4.1 Prämissen und Durchführung

Für die Evaluation der Erkennungsalgorithmen ist es erforderlich, Disaster, über die Berichte im Korpus vorliegen, zu ermitteln und relevante Tweets herauszufiltern. Dazu ist eine manuelle Analyse der 292.981.506 Tweets zu jedem der 3.007 gefundenen Extremereignisse aus der Datenbanksuche erforderlich. Jedoch liegt diese in vollumfänglicher Form außerhalb des Leistungsumfangs der Arbeit. Deswegen wurden für den weiteren Verlauf fünf Einschränkungen und Bedingungen formuliert.

So sollen aufgrund der Datenvollständigkeit, das bedeutet die Datenbanken beinhalten Einträge zwischen Januar 2011 und Oktober 2019, und der 153 Korpuszeiträume zu Amokläufen an Bildungseinrichtungen die drei Desastertypen Erdbeben, Flut und Amoklauf betrachtet werden, wobei Erdbeben mindestens eine Magnitude von 4,5 auf der Richterskala aufweisen, Fluten innerhalb eines Zeitraums des Korpus entstanden sein und bei Amokläufen mindestens fünf Personen verletzt oder getötet worden sein sollen. Zusätzlich sollen die ermittelten Einträge zu Amokläufen aus der Datenbank der GVA [2020] um die Einträge aus CNN Labs [2018] zwischen 2011 und 2014 erweitert und gleichartige Extremereignisse mit raumzeitlichen Überschneidungen aggregiert werden. Die genannten Einschränkungen liegen darin begründet, dass Erdbeben erst zwischen den Magnituden 4 und 5 auf der Richterskala als spürbar interpretiert werden und Vor- und Nachbeben in der Datenbank der USGS [2020] einzeln aufgelistet sind, die Endzeitpunkte und damit die Dauer der Fluten in der Datenbank der DFO [2020] teilweise approximiert sind und Amokläufe mit weniger als fünf Opfern Suizide, Unfälle oder Beziehungstaten darstellen können, bei denen die Öffentlichkeit nicht involviert war.

Unter Beachtung dieser Prämissen wurden zum einen 2099 Erdbeben, davon 94 Vor- und Nachbeben, 287 Fluten und 73 Amokläufe entfernt und zum anderen zwei Amokläufe hinzugefügt. Damit beinhalteten die angepassten finalen Datensätze 84 Erdbeben, 46 Fluten und 51 Amokläufe. Anschließend wurden die Zeitangaben und Ortsbezeichnungen aus den Datensätzen genutzt, um nach Meldungen von Betroffenen im Korpus zu suchen.

Tabelle 3.3: Im Korpus abgedeckte Desaster mit Zeitpunkten der ersten Tweets.

Desastertyp	Zeitpunkt (UTC)
Erdbeben	10.01.2012 18:40:20
	13.11.2016 11:03:41
	02.10.2017 05:15:52
	06.07.2019 03:17:01
	10.09.2019 05:40:09
	15.10.2019 19:42:19
Flut	31.03.2011 23:35:49
	10.06.2014 15:11:11
	09.01.2018 11:35:51
Amoklauf	24.10.2014 17:46:47
	12.06.2016 06:17:13
	02.10.2017 05:13:36
	18.05.2018 23:55:04

Tabelle 3.4: Auswahl annotierter Tweets.

Datum (UTC)	Tweet-ID	Text
2011-03-31 23:35:49	53601420519276544	„Flood in April“
2012-01-10 18:40:20	156807604620890112	„Earthquake“
2014-10-24 17:46:47	525704974417010688	„Shooting?“
2016-11-13 11:03:41	797756818457923584	„That’s an earthquake.“
2017-10-02 05:13:36	914719996852371457	„SHOOTING AT ROUTE 91“
2019-07-06 03:17:01	1147343666786131968	„Earthquake just hit... fuck...“
2019-09-10 05:40:09	1171297286627844096	„Earthquake“

Es konnten insgesamt 13 Tweets gefunden und markiert werden (Auflistung der Zeitpunkte in Tabelle 3.3 und Beispiele in Tabelle 3.4), wobei sechs Tweets auf Erdbeben, drei Tweets auf Fluten, und vier Tweets auf Amokläufe entfielen.

Zur Ermittlung wurden die Tweets des Korpus mit Hilfe von Wortlisten (Tabelle A.1) gefiltert, chronologisch geordnet und nacheinander betrachtet. Die Suche nach einer Meldung über ein Desaster startete mit dem ersten Tweet, der unmittelbar nach dem Auftreten eines Desasters aus den Datensätzen verfasst wurde und endete, sobald ein Tweet als relevant eingestuft oder keine relevante Meldung im entsprechenden Zeitraum des Korpus gefunden wurde. Im Zuge der Relevanzprüfung wurden die Tweets nach Ortsangaben durchsucht und weitere Statusmeldungen sowie Profilinformationen der Autoren ausgewertet. Letzteres konnte durch die Nutzer-ID und die erweiterte Twitter-Suchfunktion realisiert werden.

3.4.2 Beobachtungen

Während der Suche nach relevanten Meldungen wurden Auffälligkeiten im Inhalt und Aufbau von Tweets festgestellt. Es konnten Duplikate, Botnachrichten, Phishing Tweets und Berichte über Desaster, die außerhalb der Korpuszeiträume auftraten, gefunden und Ambiguitäten festgestellt werden.

So befinden sich im Korpus unter anderem Retweets und Antworten, die als neue Statusmeldungen veröffentlicht wurden. Diese Tweets sind, wenn vom Autor angegeben, durch ein vorangestelltes „RT“ oder @-Zeichen erkennbar. Jedoch existieren auch Fälle, in denen Nutzer weder den Twitter-Button für Retweets nutzten noch den Vermerk „RT“ hinzufügten. Somit enthält der Korpus trotz der Filterung während der Datenakquise redundante Tweets.

Außerdem wurden Bots gefunden, die je nach Desastertyp ein anderes Verhalten aufweisen. Bei der Suche nach Erdbeben fielen insbesondere Nachrichten des japanischen Bots @zishin3255 auf, der im Minutentakt Erdbeben mit Magnituden ab 3,0 meldet und nummerierte Updates zum Verlauf veröffentlicht.

Zusätzlich wurden Bots gefunden, die über neue Einträge im „Earthquake Catalog“ der USGS informieren, wie beispielsweise der Bot @everyEarthquake. Bei der Suche nach Fluten wurden Bots gefunden, die Berichte aus lokalen Nachrichtenseiten aufgreifen und Warnungen während Starkregen senden. Bei der Suche nach Amokläufen waren insbesondere Nachrichten des Bots @backfiretrump der Initiative States United to Prevent Gun Violence (SUPGV [2020]) präsent, der auf die Waffengewalt in den USA hinweist. Hierbei ist hervorzuheben, dass dieser Bot nicht nur über den eigenen, sondern auch über die Accounts von Nutzern Tweets versendet, sobald diese ihren Account mit der Webseite der Kampagne verlinken. Der Bot prüft Nachrichtenmeldungen und Datenbankeinträge über Amokläufe und versendet Tweets, sobald ein Amoklauf mit mindestens einem Toten gemeldet wurde. Des Weiteren wurden bei der Suche nach Amokläufen Phishing Tweets von Bots entdeckt. So sendete ein Botnetzwerk vermeintliche Eilmeldungen über einen neuen Amoklauf an einer Highschool in Colorado, obwohl sich zu diesem Zeitpunkt keine Vorfälle an einer Schule in Colorado ereigneten.

Darüber hinaus wurden Meldungen von Personen und Nachrichtenportalen gefunden, die nicht mit den Krisen aus den Datenbankeinträgen in Verbindung stehen. So befinden sich im Korpus Berichte über reale Desaster, die außerhalb der Korpuszeiträume auftraten, wie beispielsweise das Erdbeben in Haiti am 12. Januar 2010, das Erdbeben in Nepal am 25. April, die beiden Erdbeben in Nepal und Japan am 12. Mai 2015 und das Attentat auf Christina Victoria Grimmie am 10. Juni 2016. Daher sind im Korpus Tweets über Desaster enthalten, die nicht markiert wurden.

Zudem konnte beobachtet werden, dass englische Desasterbezeichnungen mehrdeutig sind. So kann beispielsweise das Wort „shooting“ eine *Schießerei*, einen *Fototermin*, oder in Verbindung mit „star“ eine *Sternschnuppe* beschreiben. Auch „flood“, „deluge“ oder „torrent“ können sowohl im Sinne von *strömendem Wasser* als auch im Sinne von *große Menge an*, beziehungsweise zur Beschreibung der Filesharing-Technik *BitTorrent*, verwendet werden. Folglich liefert die auf Keywords basierte Suche Tweets, die in keinem Zusammenhang mit Extremereignissen stehen.

Kapitel 4

Methodik

In diesem Kapitel werden die Ursprünge und Funktionsweisen der vier Erkennungsalgorithmen Keywords, Classification, Clustering und Burstiness aufgeschlüsselt, deren Reimplementierung vorgestellt und Experimente mit dem konstruierten Textkorpus beschrieben.

4.1 Grundlagen der Ereigniserkennung

Ein Ereignis ist nach Allan [2002] und Becker et al. [2011] eine zeitlich begrenzte Aktivität in der realen Welt. In Bezug auf soziale Medien erweitern Zhao and Mitra [2007] und Zhou and Chen [2013] diese Definition, indem sie ein Ereignis durch die vier Dimensionen Inhalt, Raum, Zeit und Netzwerkinformationen charakterisieren (Beispiel in Abbildung 4.1). Diese Informationen können sowohl aus dem Text als auch aus den Metadaten eines Postings gewonnen werden.

Abbildung 4.1: Dimensionen eines Tweets.

```
text: RT @Benote19: Earthquake in San Diego #San Diego #earthquake  
created_at: 1562384838237 id: 1147346253467901952 user: 55160000
```

Dimension	Ausprägung
Inhalt	Earthquake in San Diego
Raum	San Diego
Zeit	1562384838237 (6. Juli 2019 03:47:18.237 in UTC)
Netzwerkinformationen	RT @Benote19

Der *Inhalt* wird vom Autor schriftlich übermittelt, umfasst alle Informationen bezüglich des Ereignistyps und deutet teilweise oder sogar vollständig auf die beiden nachfolgenden Ebenen *Raum* und *Zeit* hin. Herausforderungen bei der Analyse des Inhalts bestehen neben der Relevanzbeurteilung insbesondere darin, Stoppwörter und Textverschmutzungen zu erkennen. Informationen zum *Raum* und zur *Zeit* können, wenn vorhanden, durch das Geotag und den Zeitstempel aus den Metadaten eines Tweets gewonnen werden. Darüber hinaus lassen sich Informationen aus dem Inhalt extrahieren, wie beispielweise Toponyme durch Named Entity Recognition. *Netzwerkinformationen* beschreiben das Verhältnis zwischen dem Autor und seinen Followern. Beispiele hierfür sind Retweets, Antworten und Verlinkungen. Dadurch ist es möglich, einerseits Tweets über gleiche Ereignisse zu bündeln und andererseits weitere Metadaten zum *Raum* und zur *Zeit* zu erhalten.

4.2 Algorithmen zur Ereigniserkennung

State-of-the-Art-Ansätze zur Erkennung von Extremereignissen können in die vier Kategorien Keywords, Classification, Clustering und Burstiness eingeteilt werden. Da keine Quelltexte aus Forschungen öffentlich verfügbar sind, wurden die Algorithmen nach Beschreibungen aus den wissenschaftlichen Arbeiten von Olteanu et al. [2014] und Temnikova et al. [2015], Sakaki et al. [2010], Kaleel and Abhari [2014] und Kleinberg [2002] nachgebildet. Im Folgenden werden diese Ansätze vorgestellt und die Nachbildungen inklusive Anpassungen beschrieben.

4.2.1 Keywords

Die Motivation für den *Keywords*-Algorithmus ergibt sich daraus, dass einerseits Disaster durch mehrere Begriffe charakterisiert werden können und andererseits Augenzeugen ihre Wahrnehmungen unterschiedlich schildern. So kann beispielsweise ein Erdbeben durch die Wörter „earthquake“ oder „tremor“, aber auch durch „ground is shaking“ sowie „seismic activity“ beschrieben werden. Aufgrund dessen erstellten Olteanu et al. [2014] und Temnikova et al. [2015] Lexika mit 380 beziehungsweise 7.200 Termini, die krisenabhängig angepasst und zum Filtern von Tweets genutzt werden können.

Der Algorithmus beginnt mit einer trivialen Suche. Listen mit Keywords dienen der Generierung einer Query, mit welcher der Korpus durchsucht wird. Nach der Suche werden die gefundenen Tweets chronologisch geordnet, wodurch die zeitliche Verteilung der Treffer analysiert werden kann. Der Algorithmus meldet immer dann ein neues Ereignis, wenn ein Schwellenwert von Treffern (Threshold) innerhalb eines Zeitintervalls erreicht wird.

4.2.2 Classification

Mit Hilfe statistischer Modelle können Daten hinsichtlich struktureller Muster klassifiziert werden (Bishop [2006]). Daraus ergibt sich die Möglichkeit, Tweets nicht nur durch Keywords, sondern auch durch Kontext auf Relevanz zu prüfen.

Diesen Ansatz wählten Sakaki et al. [2010] und analysierten Tweets über Erdbeben, definierten fünf strukturelle Features (Tabelle 4.1) und trainierten eine Support Vector Machine (SVM). Dabei nutzten die Autoren das Keyword „earthquake“ zum Filtern der Tweets und zum Ermitteln der Features. Im Detail beschreibt das erste Feature die Wörter des Tweets, das zweite die Anzahl der Wörter, das dritte die Position des Keywords im Tweet und das vierte sowie fünfte die beiden an das Keyword angrenzenden Wörter. Der nachgebildete *Classification*-Algorithmus und das erstellte Modell basieren auf diesen Features.

Tabelle 4.1: Features des Tweets „apparently, an earthquake is going on“ mit dem Keyword „earthquake“ nach Sakaki et al. [2010].

Feature	Ausprägung
(1) Wörter	apparently, an, earthquake, is, going, on
(2) Wortanzahl	6
(3) Position des Keywords	3
(4) Wort vor dem Keyword	an
(5) Wort nach dem Keyword	is

Damit die Features einheitlich und folglich vergleichbar extrahiert werden können, sind Vorverarbeitungsprozesse für den Text und Regeln für Ausnahmefälle, wie beispielsweise für den Tweet „Earthquake!“, notwendig. So entfernt der Algorithmus zunächst nach Vorbild von Wiegmann et al. [2020] Satzzeichen und ersetzt Hashtags, Zahlen, Zeitangaben, Emoticons sowie Links durch die Kennzeichen `<hashtag>`, `<number>`, `<time>`, `<emoji>` und `<url>`, da diese variieren und Vergleiche erschweren. Danach werden die Features ermittelt, wobei Ausnahmefälle wie folgt berücksichtigt werden:

1. Wenn in einem Tweet kein Keyword enthalten ist, wird der Tweet ignoriert.
2. Jeder Tweet wird um die Kennzeichen „<sot>“ und „<eot>“ am Anfang beziehungsweise am Ende erweitert.
3. Wenn ein Keyword in einem Tweet mehrfach vorhanden ist, wird nur die erste Instanz betrachtet.

Da eine SVM nur numerische Werte verarbeiten kann, ist es notwendig, das erste, vierte und fünfte Feature auf Zahlen und Vektoren abzubilden. Dazu erstellt der Algorithmus ein Wörterbuch und eine Term-Dokument-Matrix für alle Wörter beziehungsweise Tweets der Trainingsdaten.

Sobald das Modell trainiert wurde, werden Tweets nach dem Keyword durchsucht, chronologisch geordnet und klassifiziert. Der Algorithmus meldet immer dann ein neues Ereignis, wenn der Threshold für relevant klassifizierte Tweets innerhalb eines Zeitintervalls erreicht wird.

4.2.3 Clustering

Der *Clustering*-Algorithmus von Kaleel and Abhari [2014] untersucht Tweets auf Ähnlichkeiten und bündelt diese mit Hilfe von Locality-Sensitive Hashing (LSH). Die grundlegende Annahme des Ansatzes lautet, dass Tweets über Disaster dieselben Wörter beinhalten und ähnlich strukturiert sind.

Zu Beginn werden die Tweets chronologisch geordnet und analog zum *Classification*-Algorithmus vorverarbeitet, indem Satzzeichen entfernt und Hash-tags, Zahlen, Zeitangaben, Emoticons sowie Links durch Kennzeichen ersetzt werden. Anschließend werden die Tweets untereinander verglichen, wobei der Jaccard-Koeffizient (Jaccard [1908]) als Ähnlichkeitsmaß dient:

$$J(A, B) \underset{\text{Jaccard}}{=} \frac{|A \cap B|}{|A \cup B|} \quad (4.1)$$

Dabei gilt, wenn $J(A, B) = 1$, dann sind die Wörter der Tweets A und B identisch und wenn $J(A, B) = 0$, dann divergieren alle Wörter in A und B. Beispielsweise ist der Jaccard-Koeffizient für die Tweets A „Earthquake in Burma“ und B „Earthquake in Islamabad“ 0,5, da $A \cap B = \{Earthquake, in\}$ und $A \cup B = \{Earthquake, in, Burma, Islamabad\}$. Jedoch können Tweets dieselben Wörter in unterschiedlicher Reihenfolge enthalten, wodurch ungleiche Tweets fälschlich als zueinander ähnlich gewertet werden würden. Deswegen interpretiert der Algorithmus Tweets als Folgen von Wörtern, die in Teilfolgen, sogenannte Shingles, zerlegt werden können (Müller [2010]). So kann beispielsweise der Tweet „shooting at route <number>“ in die zwei k-Shingles „shooting at route“ und „at route <number>“ mit $k = 3$ eingeteilt werden. Dieses Vorgehen ermöglicht den Vergleich von Tweets unter Beibehaltung des Kontexts. Allerdings ist die Anzahl der Zeichen innerhalb der Shingles variabel, wodurch Speicher- und Rechenprozesse erschwert sind. Außerdem sind für N Tweets mindestens:

$$\binom{N}{2} = \frac{N(N-1)}{2} \approx \frac{N^2}{2} \quad (4.2)$$

Vergleiche notwendig (McCormick [2015]). Folglich sind bereits bei 1.000 Tweets 1.000.000 Vergleiche notwendig. Deswegen wird der LSH-Ansatz MinHash nach Broder [1997] angewendet, welcher Jaccard-Koeffizienten approximiert.

Der Algorithmus bildet zuerst Shingles auf numerische Werte (MinHash-Objekte) injektiv ab, ordnet diese danach n-fach zufällig an und ermittelt schließlich für jede Permutation Gruppen von Tweets, welche die ersten gemeinsamen Minhash-Objekte der jeweiligen Permutation beinhalten (Phillips [2013]). Ein Beispiel ist in Abbildung 4.2 dargestellt. Das Ergebnis ist eine Tabelle, in der ähnliche Tweets in Buckets geclustert sind.

Abbildung 4.2: Simplifizierte Darstellung des MinHash-Algorithmus mit Beispiel nach Phillips [2013].

1. Tweets mit MinHash-Objekten:

$$T_1 = \{1, 2, 5\} \quad T_2 = \{3\} \quad T_3 = \{2, 3, 4, 6\} \quad T_4 = \{1, 4, 6\}$$

2. Repräsentation der Tweets und MinHash-Objekte als Tabelle:

MinHash-Objekt	T_1	T_2	T_3	T_4
1	1	0	0	1
2	1	0	1	0
3	0	1	1	0
4	0	0	1	1
5	1	0	0	0
6	0	0	1	1

3. Zufällige Anordnung der MinHash-Objekte:

MinHash-Objekt	T_1	T_2	T_3	T_4
2	1	0	1	0
5	1	0	0	0
6	0	0	1	1
1	1	0	0	1
4	0	0	1	1
3	0	1	1	0

4. Ermitteln der ersten 1 pro Spalte:

$$m(T_1) = 2 \quad m(T_2) = 3 \quad m(T_3) = 2 \quad m(T_4) = 6$$

5. Ähnlichkeitswerte inkrementieren (T_1 und T_3) und zurück zu Schritt 3.

Die Tweets werden iterativ verarbeitet, bis der Threshold in einem Cluster erreicht ist und damit ein neues Ereignis vorliegt. Für das ausgewählte Cluster ermittelt der Algorithmus von Kaleel and Abhari [2014] das Thema durch die Bestimmung des am häufigsten enthaltenen Terminus. Jedoch kann dieses Vorgehen aufgrund von Ambiguitäten (siehe Unterabschnitt 3.4.2) in der fälschlichen Erkennung von Extremereignissen resultieren. Deswegen wurde bei der Nachbildung ein anderer Ansatz gewählt, wobei sowohl das Trainieren einer SVM nach McCreadie et al. [2016] als auch das Erstellen eines Latent Dirichlet Allocation Wahrscheinlichkeitsmodells (Blei et al. [2003]) nach Cheng and Wicks [2014] ausgeschlossen wurden, da diese außerhalb des Leistungsumfangs der Arbeit liegen. So inkludiert der nachgebildete Algorithmus das MinHash-Objekt eines festgelegten Tweets, dessen Thema bekannt ist, und ermittelt nach dem Hinzufügen eines neuen Tweets die Größe des Clusters, das den festgelegten Tweet beinhaltet.

4.2.4 Burstiness

Der *Burstiness*-Algorithmus basiert auf der Forschungsarbeit von Kleinberg [2002]. Dabei wird die zeitliche Verteilung von Tweets observiert und ungewöhnlich frequente Vorkommnisse von Begriffen, sogenannte Bursts, ermittelt. Es wird erwartet, dass durch ein neues Ereignis die Nutzung von krisenbedingten Termini abrupt ansteigt. Der nachgebildete Algorithmus verwendet die Implementierung von Poddighe [2014].

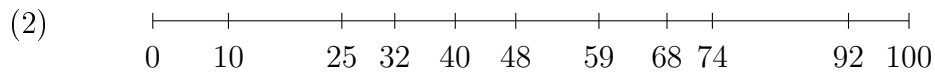
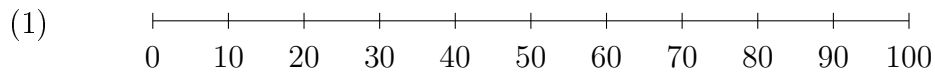
Zunächst werden Tweets nach mindestens einem Keyword durchsucht und die Zeitstempel der gefundenen Tweets (Offset) ermittelt. Anschließend wird das Offset aufsteigend geordnet und die zeitlichen Abstände nach Kleinberg [2002] zur Erstellung eines endlichen Automaten verwendet. Dieser Automat kann als $\mathcal{A}_{s,\gamma}^k$ beschrieben werden. Dabei sind γ und s Hyperparameter für die Kosten eines Zustandswechsels und k ist die Anzahl der Zustände. Gleichung 4.3 zeigt die Berechnung von k .

$$k = \lceil 1 + \log_s T + \log_s (\min_{i=1}^n x_i)^{-1} \rceil \quad (4.3)$$

In der Gleichung beschreibt x die Menge der Abstände zwischen den Elementen des Offsets und T den Abstand zwischen dem ersten und letzten Tweet. Der konstruierte Automat ermöglicht das Erkennen von Bursts durch die Berechnung von Zustandswechselkosten von Teilintervallen. Dabei gilt, dass das kleinste Teilintervall mit den höchsten Kosten dem stärksten Burst entspricht. Der Rückgabewert des Algorithmus von Poddighe [2014] ist die Menge von Teilintervallen, in denen Bursts erkannt wurden (Beispiel in Abbildung 4.3). Folglich repräsentiert die Größe der Menge die Burstiness des Offsets und der

dazugehörigen Tweets. So meldet der Algorithmus immer dann ein neues Ereignis, wenn der Threshold in der Ergebnismenge erreicht ist.

Abbildung 4.3: Darstellung von drei Offsets und Ergebnismengen $M_{1,2,3}$ des Kleinberg-Algorithmus nach Poddighe [2014] mit $s = 2$ und $\gamma = 1$. Das Offset 3 weist im Intervall $[70\ 75]$ einen Burst auf ([0 100] ist Standardelement).



$$M_1 = \{[0, 100]\} \quad M_2 = \{[0, 100]\} \quad M_3 = \{[0, 100], [70, 75]\}$$

4.3 Experimente

Zum Evaluieren der Algorithmen wurde der in Kapitel 3 vorgestellte Korpus verwendet. Das Ziel war die Erkennung der abgedeckten sechs Erdbeben, drei Fluten und vier Amokläufe.

Sowohl für den Keywords- als auch für den Burstiness-Algorithmus wurden die Listen verwendet, die in Abschnitt 3.4 erstellt wurden und in Tabelle A.1 einzusehen sind. Das ist dadurch begründet, dass in den Lexika von Olteanu et al. [2014] und Temnikova et al. [2015] auch Wortgruppen und Worte wie beispielsweise „thoughts prayers“, „massive“ und „children“ enthalten sind. Diese würden die Treffergenauigkeit beeinflussen und simultan auftretende Extremereignisse könnten nicht voneinander abgegrenzt werden. Für den Classification-Algorithmus wurden zum Erkennen von Erdbeben und Fluten Trainingsdaten von Wiegmann et al. [2020] verwendet. Da keine Trainingsdaten mit Tweets über Amokläufe vorlagen, wurden Tweets aus dem Korpus zufällig ausgewählt, nach dem Vorbild von Wiegmann et al. [2020] auf Relevanz geprüft und in einer Datei gebündelt. Als Keywords wurden die englischen Wörter „earthquake“, „flood“ und „shooting“ verwendet. Für den Clustering-Algorithmus wurden zum Auswerten der Cluster die MinHash-Objekte der Strings „earthquake“, „flood in“ und „shooting at“ verwendet. Diese Auswahl basiert auf Mustern, die in annotierten Tweets (Tabelle 3.4) gefunden wurden.

Für alle vier Algorithmen und für alle drei Desastertypen wurden Experimente mit drei verschiedenen Thresholds durchgeführt (Tabelle 4.2), sodass insgesamt 36 Programmläufe gestartet und ausgewertet wurden. Dabei sollten die Parameter zunächst bei jedem Algorithmus identisch gestaffelt werden, jedoch stellte sich bei Testversuchen heraus, dass dieser Ansatz je nach Algorithmus entweder zu viele falsche oder zu wenige korrekte Treffer lieferte.

Tabelle 4.2: Parameter für Experimente.

Algorithmus	Intervall (s)	Threshold 1	Threshold 2	Threshold 3
Keywords	10	5	10	15
Classification	10	3	5	10
Clustering	600	3	5	10
Burstiness	60	5	10	15

Jeder Programmlauf erzeugte zwei Dateien, wobei die erste alle erkannten Ereignisse sowie die Zeitpunkte der Erkennungen umfasst (Beispiel in Abbildung 4.4) und die zweite die dazugehörigen Tweets inklusive IDs und Zeitstempel beinhaltet. Dabei wurde der Erkennungszeitpunkt entweder durch den spätesten Tweet (Keywords, Classification und Clustering) oder durch die Dauer des betrachteten Intervalls (Burstiness) ermittelt. Hierbei ist anzumerken, dass alle Algorithmen nach dem Erkennen eines neuen Ereignisses für zehn Minuten alle nachfolgenden Tweets ignorierten, weil Ereignisse nicht nur während der Entstehung, sondern auch im Verlauf und darüber hinaus diskutiert werden (siehe Unterabschnitt 3.4.2), wodurch dieselben Ereignisse mehrfach hintereinander erkannt werden würden.

Abbildung 4.4: Beispiel Ausgabedatei mit gefundenen Ereignissen für Desastertyp Flut, Algorithmus Classification und Threshold 5.

```

First Tweet: 2011-04-19 14:49:27 | Event found: 2011-04-19 14:49:28
First Tweet: 2011-10-24 17:14:20 | Event found: 2011-10-24 17:14:28
First Tweet: 2012-05-25 18:00:47 | Event found: 2012-05-25 18:00:54
First Tweet: 2013-10-21 14:04:40 | Event found: 2013-10-21 14:04:42
First Tweet: 2016-06-12 09:19:20 | Event found: 2016-06-12 09:19:29
First Tweet: 2016-11-13 12:37:29 | Event found: 2016-11-13 12:37:29
First Tweet: 2018-01-09 11:41:00 | Event found: 2018-01-09 11:41:04
First Tweet: 2018-09-14 19:02:38 | Event found: 2018-09-14 19:02:47

```


Kapitel 5

Ergebnisse

Zur Evaluation wurden zunächst die Ausgabedateien mit den Erkennungszeitpunkten der Algorithmen analysiert. Alle enthaltenen Einträge wurden mit den Zeitpunkten der im Korpus abgedeckten sechs Erdbeben, drei Fluten und vier Amokläufen abgeglichen. Dabei fiel auf, dass trotz des Sprungs um zehn Minuten Ereignisse wie beispielsweise das Kaikoura-Erdbeben am 13. November 2016 bis zu 33 Mal erkannt wurden (Keywords, Threshold 5). Da diese Einträge keine fälschlich erkannten Ereignisse repräsentieren, wurden alle tagesgleichen Einträge nach der ersten Erkennung aggregiert. Die detaillierten Analyseergebnisse inklusive Recall, Precision, F1-Wert und der durchschnittlichen Erkennungszeit sind in den angehängten Tabellen B.1 (Keywords), B.2 (Classification), B.3 (Clustering) und B.4 (Burstiness) dargestellt. Zellen ohne numerische Werte implizieren, dass der Algorithmus leere Ausgabedateien lieferte beziehungsweise für die Berechnung der durchschnittlichen Erkennungszeit keine Daten vorlagen. Die im Folgenden aufgelisteten durchschnittlichen F1-, Recall- und Precisionwerte sowie Erkennungszeiten basieren auf diesen Tabellen.

Tabelle 5.1: Zusammenstellung der besten F1-Werte aller Experimente und besten Erkennungszeiten eines einzelnen Desasters.

Algorithmus	Bester F1-Wert	Beste Erkennungszeit (s)
Keywords	0,6154 (Erdbeben)	37 (Amoklauf)
Classification	0,4000 (Erdbeben)	26 (Erdbeben)
Clustering	0,9091 (Erdbeben)	38 (Erdbeben)
Burstiness	0,4211 (Erdbeben)	226 (Amoklauf)

Die niedrigsten Erkennungszeiten eines einzelnen Desasters und die höchsten F1-Werte aus allen Experimenten sind in Tabelle 5.1 dargestellt. Es lässt sich

entnehmen, dass der Clustering-Algorithmus beim Erkennen von Erdbeben den besten F1-Wert (0,9091) erzielte und der Classification-Algorithmus, ebenfalls beim Erkennen von Erdbeben, die beste Erkennungszeit eines einzelnen Desasters (26 Sekunden) erreichte. Generell fällt auf, dass alle Algorithmen Erdbeben mit den besten F1-Werten erkannten und der Desastertyp Flut weder mit dem besten F1-Wert noch der besten Erkennungszeit vertreten ist.

Tabelle 5.2: Durchschnittliche Recall-, Precision- und F1-Werte sowie Erkennungszeiten.

Algorithmus	Ø Recall	Ø Precision	Ø F1-Wert	Ø Zeit (s)
Keywords	0,6388	0,2655	0,3132	116
Classification	0,5833	0,3050	0,3050	119
Clustering (mit Flut)	0,3611	0,6667	0,4364	-
Clustering (ohne Flut)	0,5416	1,0000	0,6546	3329
Burstiness	0,5833	0,3055	0,2873	772

Die durchschnittlichen F1-, Recall- und Precisionwerte sowie Erkennungszeiten sind in Tabelle 5.2 gelistet. Für die Berechnungen wurden die höchsten F1-, Recall- und Precisionwerte und die niedrigsten Erkennungszeiten für jeden Desastertyp unabhängig vom Threshold ausgewählt (Werte in Anhang B). Aus der Tabelle geht hervor, dass sich der Clustering-Algorithmus durch den besten durchschnittlichen F1-Wert (0,4364 mit Flut, 0,6546 ohne) sowie die beste durchschnittliche Precision (0,6667 mit Flut, 1,0000 ohne) und der Keywords-Algorithmus durch den besten Recall (0,6388) sowie die beste durchschnittliche Erkennungszeit (116 Sekunden) auszeichnet. Es wird deutlich, dass der Burstiness-Algorithmus den niedrigsten durchschnittlichen F1-Wert (0,2873) und der Clustering-Algorithmus die höchste durchschnittliche Erkennungszeit (3329 Sekunden, 28-fach höher als der Keywords-Algorithmus) aufweist.

Tabelle 5.3: Ursachen für fälschlich erkannte Ereignisse.

Algorithmus	Typ	Spam	Bot	News	Ambiguität	Ongoing
Keywords mit Threshold: 10	Erdbeben	4	7	0	0	1
	Flut	4	3	0	0	0
	Amoklauf	5	0	4	1	0
Classification mit Threshold: 5	Erdbeben	6	2	3	0	1
	Flut	7	0	1	0	0
	Amoklauf	13	0	6	4	0
Clustering mit Threshold: 5	Erdbeben	0	0	0	0	0
	Flut	-	-	-	-	-
	Amoklauf	0	0	0	0	0
Burstiness mit Threshold: 10	Erdbeben	1	3	0	0	1
	Flut	2	7	0	0	0
	Amoklauf	0	1	2	5	0

Tabelle 5.4: Tweets aus fälschlich erkannten Ereignissen, jeweils ein Tweet für Erdbeben, Fluten und Amokläufe.

Algorithmus	Tweet
Keywords	„#earthquake M 2.0 - 35km SE of Cape Yakataga, Alaska“ „Yayyy flood warning on my tv. Joy.“ „You wasn’t wit me shooting in the gym!!!“
Classification	„PSN outage will cost Sony \$ <number>, earthquake bill...“ „Noses beak red flush drip worn coral schnoz flood whiff“ „To keep it a band, Tommy a bitch for shooting Angie.“
Clustering	Keine irrelevanten Tweets. Keine Daten. „shooting starsssss“
Burstiness	„1.93 earthquake occurred near Taupo, Waikato...“ „HOLLIDAY HOME RUN!!! The flood gates open...“ „Yooooooo I just seen my first shooting star !!“

Um zu ermitteln, weshalb die Algorithmen Ereignisse fälschlich erkennen, wurden die Tweet-Ausgabedateien für jeden Algorithmus ausgewertet. Die entnommenen Ursachen sind in Tabelle 5.3 zusammengefasst und ausgewählte Tweets aus fälschlich erkannten Ereignissen können in Tabelle 5.4 eingesehen werden.

Dabei wird ein erkanntes Ereignis unabhängig vom Inhalt als *Spam* bezeichnet, wenn mindestens 80 % der enthaltenen Tweets entweder identisch sind oder sich in maximal drei Wörtern unterscheiden (Beispiel in Abbildung C.1). Die Einteilung in die anderen vier Rubriken erfolgt nach dem Mehrheitsprinzip. Hierbei werden Ereignisse als *Bot* und *News* klassifiziert, wenn die Tweets von Bots (Beispiel in Abbildung C.2) beziehungsweise von Nachrichtenportalen stammen (Beispiel in Abbildung C.3) und Ereignisse beschreiben, die durch die Prämissen der Annotation ausgeschlossen wurden (siehe Unterabschnitt 3.4.1), als *Ambiguität*, wenn die Tweets krisenunabhängige Themen beinhalten (Beispiel in Abbildung C.4) und als *Ongoing*, wenn sich der Inhalt der Tweets auf reale Disaster bezieht, die nicht während der Korpuszeiträume entstanden (Beispiel in Abbildung C.5). Es fällt auf, dass Parallelen zu den Beobachtungen aus Unterabschnitt 3.4.2 bestehen.

Da der Keywords-Algorithmus Tweets unabhängig vom Kontext verarbeitet, werden neben erkenntnisrelevanten Tweets zum einen Spam (13 Ereignisse) und zum anderen Meldungen von Bots über Erdbeben mit Magnituden kleiner als 4,5 auf der Richterskala, automatisierte Flutwarnmeldungen nach Starkregen und Nachrichten über Amokläufe mit weniger als fünf Opfern (14 Ereignisse) einbezogen. Außerdem wird das Erdbeben in Nepal am 12. Mai 2015 erkannt, weil der Korpuszeitraum vom 13. Mai 2015 Nachrichtmeldungen über entstandene Schäden und Folgen beinhaltet. Auch der Classification-Algorithmus wertet Spam (26 Ereignisse) und Nachrichtmeldungen (7 Ereignisse) als relevant und erkennt das Erdbeben in Nepal, schließt jedoch Tweets von Bots (2 Ereignisse) nahezu vollständig aus. Der Grund dafür liegt im Trainingsdatensatz, der Nachrichten von Bots als irrelevant wertet. Der Clustering-Algorithmus selektiert vergleichsweise stark und schließt ausnahmslos Spam und alle Meldungen von Bots sowie Nachrichtenportalen aus. Dass das Erdbeben in Nepal nicht erkannt wird, liegt daran, dass die im Korpus enthaltenen Berichte hauptsächlich von Nachrichtenseiten stammen. Außerdem fällt auf, dass der Clustering-Algorithmus bei der Erkennung von Fluten keine Ergebnisse liefert, weshalb anzunehmen ist, dass sich Tweets über Fluten zum einen untereinander und zum anderen gegenüber Tweets über Erdbeben und Amokläufe hinsichtlich des Kontextes und der Struktur unterscheiden. Der Burstiness-Algorithmus verhält sich ähnlich wie der Keywords-Algorithmus, da nicht der Kontext der Tweets bewertet wird. So sind in der Ergebnismenge Spam (3 Ereignisse), Meldungen von Bots (11 Ereignisse) und Nachrichtenportalen (2 Ereignisse), Tweets über krisenunabhängige Themen (5 Ereignisse) und das Erdbeben in Nepal enthalten.

Tabelle 5.5: Erkannte (×) und nicht erkannte (-) Ereignisse.

	Datum	Keywords	Classification	Clustering	Burstiness
Erdbeben	10.01.2012	-	-	×	-
	13.11.2016	×	×	×	×
	02.10.2017	×	-	-	×
	06.07.2019	×	×	×	×
	10.09.2019	×	×	×	-
	15.10.2019	×	×	×	×
Flut	31.03.2011	-	-	-	-
	10.06.2014	-	-	-	-
	09.01.2018	×	×	-	×
Amoklauf	24.10.2014	-	-	-	-
	12.06.2016	×	×	×	×
	02.10.2017	×	×	-	×
	18.05.2018	×	×	-	×

Ein Abgleich der erkannten realen Disaster mit den im Korpus abgedeckten Extremereignissen ist in Tabelle 5.5 dargestellt. Es wird deutlich, dass der Keywords-Algorithmus die meisten Ereignisse erkennen konnte (9), gefolgt von den Classification- und Burstiness-Algorithmen (jeweils 8) und dem Clustering-Algorithmus (6). Außerdem zeigt die Tabelle, dass drei Instanzen von keinem Algorithmus erkannt wurden. Dabei handelt es sich um den Amoklauf an der Marysville Pilchuck High School am 24. Oktober 2014 und die beiden Fluten am 31. März 2011 und am 10. Juni 2014. Daher ist davon auszugehen, dass entweder zu wenige Tweets im Korpuszeitraum vorlagen oder die betrachteten Intervalle zu kurz waren. Es ist anzumerken, dass dieser Umstand in Kombination mit dem im Korpus abgedeckten Erdbeben in Nepal die Auswertung und die berechneten F1-, Recall- und Precisionwerte beeinflusst, jedoch Vergleiche zwischen den Algorithmen zur Orientierung für künftige Forschung erlaubt.

Tabelle 5.6: Durchschnittliche Laufzeiten aus zehn Programmläufen unter Verarbeitung von 235.188 Tweets des Korpus (gefiltert mit Wortlisten, entspricht Anteil von 0,08 %) und Skalierung auf täglichen Twitterstream.

Algorithmus	Laufzeit (in Sekunden)		
	Gesamt	10.000 Tweets	Twitterstream
Keywords	0,24	0,01	0,41
Classification (mit Training)	86,77	3,70	147,58
Classification (ohne Training)	84,50	3,60	143,71
Clustering	66,80	2,84	113,61
Burstiness	6,90	0,29	11,74

Die ermittelten durchschnittlichen Laufzeiten der Algorithmen sind in Tabelle 5.6 dargestellt. Zur Berechnung wurden zehn Programmläufe mit 235.188 Tweets zum Erkennen von Erdbeben durchgeführt. Die Tweets wurden zuvor mit der Wortliste für Erdbeben (siehe Tabelle A.1) aus dem Korpus herausgefiltert und spiegeln 0,08 % des Datensatzes wider. Der Tabelle lässt sich entnehmen, dass der Keywords- (0,24 Sekunden) und der Burstiness-Algorithmus (6,90) sehr schnell sind. Da sowohl der Classification- (86,77 Sekunden mit Training, 84,50 Sekunden ohne) als auch der Clustering-Algorithmus (66,80 Sekunden) Vorverarbeitungsprozesse benötigen und für jeden einzelnen Tweet Berechnungen durchführen, beanspruchen diese beiden Algorithmen mehr Rechenzeit. Wird der Anteil (0,08 %) auf den Twitterstream mit täglich 500 Millionen Tweets (Twitter Inc. Business [2019]) übertragen, ergeben sich 400.000 Tweets pro Tag. Es fällt auf, dass bei dieser Skalierung die Verzögerungen für alle Algorithmen vergleichsweise gering ausfallen, denn die niedrigste Laufzeit beträgt 0,41 Sekunden (Keywords-Algorithmus) und die höchste 147,58 Sekunden (Classification-Algorithmus inklusive Training).

Aus den beschriebenen Erkenntnissen können Hypothesen zur Leistungsfähigkeit und Übertragbarkeit der Erkennungsalgorithmen aufgestellt werden. So sind der Keywords-, der Classification- und der Burstiness-Algorithmus übertragbar, da alle drei Algorithmen sowohl Erdbeben und Fluten als auch Amokläufe erkennen. Allerdings weist von diesen Algorithmen der Burstiness-Algorithmus die höchste durchschnittliche Erkennungszeit und den niedrigsten F1-Wert auf, da der Algorithmus voraussetzt, dass eine zuvor nicht abschätzbare Mindestanzahl von Nutzern gleichzeitig von einem Disaster betroffen ist oder darüber berichtet. Folglich ist davon auszugehen, dass der Algorithmus zwar übertragbar, jedoch nicht zur universellen Erkennung von Extremereignissen geeignet ist, denn Naturkatastrophen wie beispielsweise Lawinen würden vermutlich unerkannt bleiben. Ein Vergleich zwischen dem Keywords- und Classification-Algorithmus zeigt, dass die durchschnittlichen Erkennungszei-

ten und F1-Werte ähnlich sind, sich die Algorithmen jedoch in den Ursachen für fälschlich erkannte Ereignisse unterscheiden. So meldet der Keywords-Algorithmus mehr Ereignisse aufgrund von Bots und der Classification-Algorithmus mehr Ereignisse wegen Spam und Nachrichtenmeldungen. Grund dafür sind die Intervalle, die von den Algorithmen betrachtet werden und nicht standardisiert werden können (siehe Abschnitt 4.3). Der Clustering-Algorithmus weist zwar bei der Erkennung von Erdbeben den höchsten F1-Wert auf, lässt sich aber nicht auf Fluten und nur ansatzweise auf Amokläufe übertragen. Darüber hinaus ist anzunehmen, dass alle Algorithmen zur Ereigniserkennung in Echtzeit über den gefilterten Twitterstream angewendet werden können, wobei der Classification-Algorithmus die höchste Verzögerung aufweist.

Kapitel 6

Diskussion

Die Ergebnisse zeigen, dass sowohl der Keywords- als auch der Classification-Algorithmus zur Erkennung von Extremereignissen über Twitter nutzbar und für die weitere Forschung geeignet sind. Außerdem besteht die Möglichkeit, die Precision und damit die F1-Werte der Algorithmen durch Textverarbeitung zu erhöhen. Da der Classification-Algorithmus auf Keywords basiert und zusätzlich den Kontext einbezieht, birgt dieser für die weitere Verarbeitung das größte Potenzial. Verbesserungsvorschläge für den Classification-Algorithmus umfassen die semantische Analyse des Tweet-Inhalts, die Anpassung der Trainingsdaten und das Überprüfen erkannter Ereignisse.

So ist es zunächst ratsam, die durch den Algorithmus betrachteten Keywords zu erweitern. Der nachgebildete Prototyp betrachtet zwar die Wörter „earthquake“, „flood“ und „shooting“, ignoriert aber Tweets mit Synonymen und Umschreibungen. Hier ist eine Erweiterung um die in Tabelle A.1 genannten Begriffe zu empfehlen. Zusätzlich wäre es sinnvoll, weitere Sprachen aus Risikogebieten hinzuzufügen, wie beispielweise in Bezug auf Erdbeben Japanisch und Indonesisch. Außerdem wäre es angebracht, vor der Ermittlung der Features Kookkurrenzen zu betrachten, Stammformreduktionen durchzuführen und Stoppwörter zu entfernen. So könnten zum einen Mehrdeutigkeiten wie „shooting star“ vermieden und zum anderen der Umfang des Wörterbuchs sowie der Term-Dokument-Matrix verringert werden. Dies würde in einer erhöhten Vergleichbarkeit und damit in einer verbesserten Relevanzprüfung der SVM resultieren. Zusätzlich wäre es sinnvoll, auch Eigennamen wie Anthroponyme oder Toponyme und Medien durch einheitliche Kennzeichen zu ersetzen.

Darüber hinaus könnte die Anpassung des Trainingsdatensatzes die Precision des Algorithmus erhöhen. Vorschläge umfassen die Erweiterung um Bot- und Nachrichtenmeldungen sowie um Tweets von korrekt erkannten Ereignissen. Diese könnten durch die Analyse der Ausgabedateien aller Algorithmen gewonnen werden.

Die häufigste Ursache für fälschlich erkannte Ereignisse ist Spam. Da sich Spam jedoch konstant ändert und somit nicht dauerhaft und vollständig durch die Trainingsdaten abgedeckt werden kann, ist es empfehlenswert, die erkannten Ereignisse zu überprüfen. So ist es sinnvoll, vor der Bestätigung eines neuen Ereignisses die enthaltenen Tweets auf Duplikate und Ähnlichkeiten zu prüfen. Hierfür könnten sich Vergleiche zwischen den Vektoren des ersten Features (Wörter) oder die Anwendung einer Hashfunktion wie MD5 (Rivest [1992]) oder MinHash (Broder [1997]) eignen. Da diese Prozesse die Rechenzeit erhöhen, ist es ratsam, nur Ereignisse mit Verdacht zu überprüfen, beispielsweise, wenn alle Tweets die gleiche Position des Keywords oder einen Link aufweisen. Weiterhin ist es notwendig, die gefundenen Ereignisse nicht nur zu erkennen, sondern auch zu speichern beziehungsweise nachzuverfolgen. So würde der im Prototyp implementierte Sprung entfallen und Ereignisse wie das Erdbeben in Nepal würden zu einem späteren Zeitpunkt nicht erneut erkannt werden. Hierbei wäre es auch sinnvoll, den Ort des Geschehens zu ermitteln, denn dieser würde zum einen den Nutzen erweitern und zum anderen die Erkennung von simultan aktiven Extremereignissen erleichtern, da diese besser auseinandergehalten werden könnten. Treten Ereignisse gleichzeitig auf, kann der Algorithmus diese nicht voneinander unterscheiden, da die Tweets im Bündel ein Ereignis repräsentieren. Zwar wäre eine Unterscheidung der enthaltenen Sprachen oder Toponyme denkbar, jedoch garantiert dieses Vorgehen die separate Erkennung nicht.

Der Classification-Algorithmus lässt sich auf andere Desastertypen übertragen, erkennt neue Ereignisse in durchschnittlich zwei Minuten nach dem ersten Tweet und birgt Potenzial zur Verminderung fälschlich erkannter Treffer. Die ermittelte Laufzeit zur Verarbeitung des gefilterten täglichen Twitterstreams (143,71 Sekunden) deutet darauf hin, dass der Algorithmus zur Echtzeiterkennung von natürlichen und anthropogenen Katastrophen genutzt werden kann. Dementsprechend könnte sich der Algorithmus dazu eignen, Informationen aus Krisengebieten zu erhalten und Warnsysteme zu erstellen, um Menschen zu informieren, Schäden abzuschätzen und Hilfeleistungen zu initialisieren. Aufgrund der gezeigten Übertragbarkeit auf Erdbeben, Fluten und Amokläufe ist zu vermuten, dass der Algorithmus zur Erkennung von weiteren Extremereignissen wie beispielsweise Vulkanausbrüchen, Waldbränden, Hagelstürmen und Schlammlawinen, aber auch terroristischen Anschlägen und Epidemien, nutzbar ist.

Kapitel 7

Zusammenfassung

In dieser Arbeit wurde ein Textkorpus mit 292.981.506 Tweets manuell annotiert und damit ein neuer Standard für die einheitliche Evaluation von Algorithmen zur Erkennung von Extremereignissen geschaffen. Das vorgestellte Evaluierungsschema gewährt Einblicke in die Übertragbarkeit der Algorithmen und ermöglicht das Aufstellen von Verbesserungsvorschlägen. Dazu wurden die vier Ansätze Keywords, Classification, Clustering und Burstiness nachgebildet und deren Leistungsfähigkeit ausgewertet. Der Vergleich zwischen den Algorithmen zeigt, dass sowohl der Keywords- (F1-Wert 0,3132) und der Classification- (F1-Wert 0,3050) als auch der Burstiness-Algorithmus (F1-Wert 0,2873) übertragbar sind. Der Clustering-Algorithmus weist zwar den besten durchschnittlichen F1-Wert (0,4364) auf, erkennt aber keine Fluten. Die besten Erkennungszeiten erzielten der Keywords- (116 Sekunden) und der Classification-Algorithmus (119 Sekunden), während die schlechtesten Zeiten auf den Burstiness-Algorithmus (772 Sekunden) und den Clustering-Algorithmus (3329 Sekunden, ohne Fluten) entfielen. Die Analyse der erkannten Ereignisse zeigt, dass der Keywords-Algorithmus 69 %, der Classification- sowie der Burstiness-Algorithmus jeweils 62 % und der Clustering-Algorithmus 46 % der abgedeckten Erdbeben, Fluten und Amokläufe erkannten. Dabei entfielen die drei primären Ursachen für fälschlich erkannte Ereignisse auf Spam (45 %), Tweets von Bots (25 %) und Nachrichtenmeldungen (17 %). Die Auswertung zeigt, dass der Classification-Algorithmus für die weitere Forschung das größte Potenzial birgt, da dieser auch den Kontext eines Tweets zur Relevanzprüfung einbezieht. Zukünftige Aufgaben umfassen die Verbesserung des Algorithmus durch die Reduzierung fälschlich erkannter Ereignisse, die Geolokalisierung betroffener Twitter-Nutzer und die Ereigniserkennung in Echtzeit über den Twitterstream.

Literaturverzeichnis

- Asian Disaster Reduction Centre ADRC. GLIDENumber Database, 2020. URL <https://glidenummer.net/glide/public/search/search.jsp>. Abgerufen am 09.01.2020. 3.3
- James Allan. *Topic Detection and Tracking: Event-Based Information Organization*. Springer US, 2002. doi: 10.1007/978-1-4615-0933-2. 4.1
- James Allan, Ron Papka, and Victor Lavrenko. On-line New Event Detection and Tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45. Association for Computing Machinery, 08 1998. doi: 10.1145/290941.290954. 2
- Samar M. Alqhtani, Suhuai Luo, and Brian Regan. Fusing Text and Image for Event Detection in Twitter. *The International Journal of Multimedia & Its Applications*, 7:27–35, 02 2015. doi: 10.5121/ijma.2015.7103. 2
- Hila Becker, Mor Naaman, and Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, pages 438–441, 01 2011. 2, 4.1
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 08 2006. ISBN 9780387310732. 4.2.2
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 03 2003. 4.2.3
- Andrei Z. Broder. On the Resemblance and Containment of Documents. In *Proceedings of the International Conference on Compression and Complexity of Sequences*, pages 21–29, 06 1997. doi: 10.1109/SEQUEN.1997.666900. 2, 4.2.3, 6

- Hongyun Cai, Yang Yang, Xuefei Li, and Zi Huang. What are Popular: Exploring Twitter Features for Event Detection, Tracking and Visualization. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 89–98. Association for Computing Machinery, 10 2015. doi: 10.1145/2733373.2806236. 2
- Tao Cheng and Thomas Wicks. Event Detection using Twitter: A Spatio-Temporal Approach. *PloS one*, 9:1–10, 06 2014. doi: 10.1371/journal.pone.0097807. 4.2.3
- CNN Labs. CNN School Shooting Data, 2018. URL <https://github.com/cnnlabs/cnn-school-shooting-data/blob/master/dataset.csv>. Abgerufen am 19.02.2020. 3.2, 3.4.1
- Mário Cordeiro and João Gama. *Online Social Networks Event Detection: A Survey*, volume 9580, pages 1–41. Springer International Publishing, 07 2016. doi: 10.1007/978-3-319-41706-6_1. 2
- Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. 'OMG, from here, I can see the flames!': a use case of mining Location Based Social Networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pages 73–80. Association for Computing Machinery, 11 2009. doi: 10.1145/1629890.1629907. 2
- G.R. Brakenridge DFO. Global Active Archive of Large Flood Events, 2020. URL <http://floodobservatory.colorado.edu/Archives/index.html>. Abgerufen am 12.01.2020. 3.3, 3.4.1
- Paul Earle, Daniel Bowden, and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of geophysics*, 54:708–715, 01 2012. doi: 10.4401/ag-5364. 2
- European Severe Storms Laboratory ESSL. European Severe Weather Database, 2020. URL <https://www.eswd.eu>. Abgerufen am 09.01.2020. 3.3
- Atefeh Farzindar and Wael Khreich. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, 31:132–164, 09 2013. doi: 10.1111/coin.12017. 2
- Wei Feng, Chao Zhang, Wei Zhang, Jiawei Han, Jianyong Wang, Charu Aggarwal, and Jianbin Huang. STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration over the Twitter Stream. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1561–1572, 04 2015. doi: 10.1109/ICDE.2015.7113425. 2

- Davies, Richard and Behrend, Jana and Hill, Ed FloodList. FloodList, 2020. URL <http://floodlist.com>. Abgerufen am 25.02.2020. 3.3
- Gabriel P. C. Fung, Jeffrey X. Yu, Philip S. Yu, and Hongjun Lu. Parameter Free Bursty Events Detection in Text Streams. In *Proceedings of the 31st International Conference on Very Large Data Bases*, pages 181–192. VLDB Endowment, 08 2005. ISBN 1595931546. 2
- Daniel A. Gruber, Ryan E. Smerek, Melissa C. Thomas-Hunt, and Erika H. James. The real-time power of Twitter: Crisis management and leadership in an age of social media. *Business Horizons*, 58:163–172, 01 2015. doi: 10.1016/j.bushor.2014.10.006. 1
- Gun Violence Archive GVA. Gun Violence Archive, 2020. URL <https://www.gunviolencearchive.org>. Abgerufen am 15.01.2020. 3.3, 3.4.1
- Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. Breaking News on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2754. Association for Computing Machinery, 05 2012. doi: 10.1145/2207676.2208672. 1
- Paul Jaccard. Nouvelles Recherches Sur La Distribution Florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 44:223–270, 01 1908. doi: 10.5169/seals-268384. 4.2.3
- Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. *Why We Twitter: An Analysis of a Microblogging Community*, pages 118–138. Springer Berlin Heidelberg, 2009. ISBN 9783642005282. 1
- Shakirabanu Kaleel and Abdolreza Abhari. Cluster-Discovery of Twitter Messages for Event Detection and Trending. *Journal of Computational Science*, 6:47–57, 11 2014. doi: 10.1016/j.jocs.2014.11.004. 1, 4.2, 4.2.3, 4.2.3
- Jon Kleinberg. Bursty and Hierarchical Structure in Streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 7, pages 91–101. Association for Computing Machinery, 07 2002. doi: 10.1145/775047.775061. 1, 2, 4.2, 4.2.4
- Pranam Kolari, Akshay Java, Timothy W. Finin, Tim Oates, and Anupam Joshi. Detecting Spam Blogs: A Machine Learning Approach. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1351–1356. Association for the Advancement of Artificial Intelligence, 07 2006. 2

- Rui Li, Kin Lei, Ravi Khadiwala, and Kevin Chang. TEDAS: A Twitter-based Event Detection and Analysis System. In *IEEE 28th International Conference on Data Engineering*, pages 1273–1276, 04 2012. doi: 10.1109/ICDE.2012.125. 2
- Chris McCormick. MinHash Tutorial with Python Code, 06 2015. URL <https://mccormickml.com/2015/06/12/minhash-tutorial-with-python-code>. Abgerufen am 12.03.2020. 4.2.3
- Richard McCreadie, Craig Macdonald, and Iadh Ounis. EAIMS: Emergency Analysis Identification and Management System. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1101–1104. Association for Computing Machinery, 07 2016. doi: 10.1145/2911451.2911460. 2, 4.2.3
- Cristian Müller. Erkennung naher Duplikate von Nachrichten-Dokumenten aus dem Web, 04 2010. 4.2.3
- National Aeronautics and Space Administration NASA. Global Landslide Catalog, 2015. URL <https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog/h9d8-neg4>. Abgerufen am 12.01.2020. 3.3
- Nic Newman. Mainstream media and the distribution of news in the age of social media. Technical report, University of Oxford, 09 2011. 1
- Storm Prediction Center NOAA. Severe Weather Database, 2019. URL <https://www.spc.noaa.gov/wcm/#data>. Abgerufen am 12.01.2020. 3.3
- World Data Service NOAA, National Geophysical Data Center. The Significant Volcanic Eruption Database, 2020. URL <https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ngdc.mgg.hazards:G10147>. Abgerufen am 12.01.2020. 3.3
- National Oceanic and Atmospheric Administration NWS. National Weather Service, 2020. URL <https://www.weather.gov>. Abgerufen am 27.02.2020. 3.3
- Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisis-Lex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*, pages 376–385, 01 2014. 1, 2, 4.2, 4.2.1, 4.3

- Miles Osborne, Sean Moran, Richard McCreadie, Alexander von Lunen, Martin Sykora, Amparo Cano Basave, Neil Ireson, Craig MacDonald, Iadh Ounis, Thomas Jackson, Fabio Ciravegna, and Ann O'Brien. Real-Time Detection, Tracking and Monitoring of Automatically Discovered Events in Social Media. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42. Association for Computational Linguistics, 06 2014. doi: 10.3115/v1/P14-5007. 2
- Jeff M. Phillips. Min Hashing, 2013. URL <https://www.cs.utah.edu/~jeffp/teaching/cs5955/L5-Minhash.pdf>. Abgerufen am 12.03.2020. 4.2.3, 4.2
- Barbara Poblete, Jheser Guzmán, Jazmine Maldonado, and Felipe Tobar. Robust Detection of Extreme Events Using Twitter: Worldwide Earthquake Monitoring. *IEEE Transactions on Multimedia*, 20:2551–2561, 07 2018. doi: 10.1109/TMM.2018.2855107. 2
- Renzo Poddighe. pybursts 0.1.1, 2014. URL <https://pypi.org/project/pybursts/0.1.1>. Abgerufen am 15.11.2019. 4.2.4, 4.2.4, 4.3
- Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 103–110. Association for Computing Machinery, 07 2007. doi: 10.1145/1277741.1277762. 2
- Ronald L. Rivest. The MD5 Message-digest Algorithm. *Internet Engineering Task Force*, 04 1992. doi: 10.17487/RFC1321. 6
- Zafar Saeed, Rabeeh A. Abbasi, Onaiza Maqbool, Abida Sadaf, Imran Razzak, Ali Daud, Naif R. Aljohani, and Guandong Xu. What's Happening Around the World? A Survey and Framework on Event Detection Techniques on Twitter. *Journal of Grid Computing*, pages 279–312, 06 2019. doi: 10.1007/s10723-019-09482-2. 2
- Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. Association for Computing Machinery, 2010. doi: 10.1145/1772690.1772777. 1, 2, 4.2, 4.2.2, 4.1

- Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. Association for Computing Machinery, 11 2009. doi: 10.1145/1653771.1653781. 2
- Koichi Sato, Junbo Wang, and Zixue Cheng. Credibility Evaluation of Twitter-based Event Detection by a Mixing Analysis of Heterogeneous Data. *IEEE Access*, 7:1095–1106, 12 2018. doi: 10.1109/ACCESS.2018.2886312. 2
- Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. Event Detection and Tracking in Social Streams. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, pages 311–314, 05 2009. 2
- States United to Prevent Gun Violence SUPGV. States United to Prevent Gun Violence, 2020. URL <https://supgv.org/backfiretrump-makes-the-observer>. Abgerufen am 26.02.2020. 3.4.2
- Kaneko Takamu, Nga Do Hang, and Keiji Yanai. Twitter Event Photo Detection Using both Geotagged Tweets and Non-geotagged Photo Tweets. In *Advances in Multimedia Information Processing – PCM 2015*, pages 128–138, 09 2015. doi: 10.1007/978-3-319-24078-7_13. 2
- Irina Temnikova, Carlos Castillo, and Sarah Vieweg. EMTerms 1.0: A Terminological Resource for Crisis Tweets. In *ISCRAM 2015 Conference Proceedings - 12th International Conference on Information Systems for Crisis Response and Management*, pages 147–157. Information Systems for Crisis Response and Management, ISCRAM, 2015. ISBN 9788271177881. 2, 4.2, 4.2.1, 4.3
- Twitter Inc. Selected Company Metrics and Financials, 2019. URL https://s22.q4cdn.com/826641620/files/doc_financials/2019/q3/Q3-2019-Selected-Financials-and-Metrics.pdf. Abgerufen am 04.02.2020. 1
- Twitter Inc. Tweet objects, 2020. URL <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json>. Abgerufen am 20.02.2020. 3.1
- Twitter Inc. Business. Twitter Business Landing Page, 2019. URL <https://business.twitter.com>. Abgerufen am 05.10.2019. 1, 5
- United States Geological Survey USGS. Earthquake Catalog, 2020. URL <https://earthquake.usgs.gov/earthquakes/search>. Abgerufen am 12.01.2020. 3.3, 3.4.1

- Andreas Weiler, Michael Grossniklaus, and Marc H. Scholl. An Evaluation of the Run-time and Task-based Performance of Event Detection Techniques for Twitter. *Information Systems*, 62:207–219, 12 2016. doi: 10.1016/j.is.2016.01.003. 2
- Jianshu Weng and Bu-Sung Lee. Event Detection in Twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 401–408, 07 2011. 2
- Matti Wiegmann, Jens Kersten, Frederike Klan, Martin Potthast, and Benno Stein. Analysis of Filtering Models for Disaster-Related Tweets. unveröffentlicht, 05 2020. 4.2.2, 4.3
- Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, Tom Pierce, Brian T. Archibald, and Xin Liu. Learning Approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems and their Applications*, 14:32–43, 07 1999. doi: 10.1109/5254.784083. 2
- Yiming Yang, Tom Ault, Thomas Pierce, and Charles W. Lattimer. Improving Text Categorization Methods for Event Tracking. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–72. Association for Computing Machinery, 05 2000. doi: 10.1145/345508.345550. 2
- Qiankun Zhao and Prasenjit Mitra. Event Detection and Visualization for Social Text Streams. In *ICWSM 2007 - International Conference on Weblogs and Social Media*, 03 2007. 2, 4.1
- Xiangmin Zhou and Lei Chen. Event detection over twitter social media streams. *The VLDB Journal*, 23:381–400, 07 2013. doi: 10.1007/s00778-013-0320-3. 4.1

Anhang A

Wortlisten

Tabelle A.1: Wortlisten zur Filterung.

Erdbeben	Flut	Amoklauf
earthquake	deluge	amok
earthquakes	flood	mass murder
epicentre	high tide	massacre
ground is shaking	inundation	rampage
ground shaking	swamping	shooting
magnitude	torrent	
seismic activity		
tremor		

Anhang B

Auswertung der Erkennungsalgorithmen

Tabelle B.1: Auswertung des Keywords-Algorithmus (Intervall: 10 Sekunden).

Typ	Threshold	Recall	Precision	F1-Wert	Ø Zeit (s)
Erdbeben	5	0,8333	0,1724	0,2857	204
	10	0,6667	0,5714	0,6154	193
	15	0,5000	0,6000	0,5455	275
Flut	5	0,3333	0,0417	0,0741	105
	10	0,0000	0,0000	0,0000	-
	15	0,0000	0,0000	0,0000	-
Amoklauf	5	0,7500	0,0577	0,1071	201
	10	0,5000	0,1667	0,2500	1191
	15	0,2500	0,2500	0,2500	1855

Tabelle B.2: Auswertung des Classification-Algorithmus (Intervall: 10 Sekunden).

Typ	Threshold	Recall	Precision	F1-Wert	Ø Zeit (s)
Erdbeben	3	0,6667	0,1905	0,2963	194
	5	0,5000	0,2500	0,3333	63
	10	0,3333	0,5000	0,4000	179
Flut	3	0,3333	0,0476	0,0833	71
	5	0,3333	0,1250	0,1818	313
	10	0,0000	0,0000	0,0000	-
Amoklauf	3	0,7500	0,0536	0,1000	224
	5	0,5000	0,1111	0,1818	383
	10	0,2500	0,5000	0,3333	5803

Tabelle B.3: Auswertung des Clustering-Algorithmus (Intervall: 600 Sekunden).

Typ	Threshold	Recall	Precision	F1-Wert	Ø Zeit (s)
Erdbeben	3	0,8333	0,8333	0,8333	85
	5	0,8333	1,0000	0,9091	126
	10	0,6667	1,0000	0,8000	120
Flut	3	-	-	-	-
	5	-	-	-	-
	10	-	-	-	-
Amoklauf	3	0,2500	0,1429	0,1818	6572
	5	0,2500	1,0000	0,4000	6699
	10	-	-	-	-

Tabelle B.4: Auswertung des Burstiness-Algorithmus (Intervall: 60 Sekunden).

Typ	Threshold	Recall	Precision	F1-Wert	Ø Zeit (s)
Erdbeben	5	0,6667	0,3077	0,4211	387
	10	0,3333	0,2857	0,3077	282
	15	0,3333	0,5000	0,4000	282
Flut	5	0,3333	0,0833	0,1333	1660
	10	0,0000	0,0000	0,0000	-
	15	0,0000	0,0000	0,0000	-
Amoklauf	5	0,7500	0,1304	0,2222	373
	10	0,5000	0,2222	0,3077	3037
	15	0,2500	0,3333	0,2857	5166

Anhang C

Fälschlich erkannte Ereignisse

Abbildung C.1: Fälschlich erkanntes Ereignis (Flut) aufgrund von Spam. Ausschnitt Ausgabedatei mit Tweets, Burstiness-Algorithmus. Vollständiger Burst umfasst 153 nahezu identische Einträge innerhalb von 60 Sekunden.

```
"00": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/BGtKZJZWLs"
"01": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/ZTycwbQXna"
"02": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/sc8QFV3YIP"
"03": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/SluofdtisC"
"04": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/Xj3o9OgepA"
"05": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/BzkPisvQH28"
"06": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/5wJHHh1Xt"
"07": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/UMFyhgScVv"
"08": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/AwB9zmatwL"
"09": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/CH12fS21kg"
"10": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/mCTbn7Y1cQ"
"11": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/4lgeemoWQn"
"12": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/XjvTJEDRgg"
"13": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/h4ZK50b9tL"
"14": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/EgRntv5n03"
"15": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/wTXWQe90W1"
"16": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/9C4tFelicJc"
"17": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/hwpG4xUIDB"
"18": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/WtqcFrOWdB"
"19": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/B07Z4GvWFZ"
"20": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/oBZM4wdvQH"
"21": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/50cgcuSvjl"
"22": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/958mxmLQj4"
"23": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/wM7gloem4y"
"24": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/iaRtbAhoJ2"
"25": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/rbTIYU2KaI"
"26": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/Dj6Tmq3vEx"
"27": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/VvPuat5Uuz"
"28": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/zHssQnZdW9"
"29": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/104Bx9K1h8"
"30": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/TpxDfT5JYD"
"31": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/XQcgEU6PB5"
"32": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/4oi3UVBwK7"
"33": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/8wHkPWL4b2"
"34": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/ckYZYQvT9X"
"35": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/kXhxXfR2GT"
"36": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/1JNV21lWRd"
"37": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/KbDTU1D1lU"
"38": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/FjXnHsZr9s"
"39": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/MZct3lFKqK"
"40": "Dreamlike images flood your consciousness today, challenging y... More for Pisces http://t.co/FZRnqfAuCU"
```

Abbildung C.2: Fälschlich erkanntes Ereignis (Erdbeben) aufgrund eines Bots. Ausschnitt Ausgabedatei mit Tweets, Classification-Algorithmus.

"00": "Minor earthquake - Offshore Coquimbo, Chile on June 11, 2016 #tab <https://t.co/WyzH1uPtO9>",
 "01": "Minor earthquake - Offshore Coquimbo, Chile on June 11, 2016 #tab <https://t.co/mUePrHIJEV>",
 "02": "Moderate earthquake - San Juan, Argentina on June 12, 2016 #tab <https://t.co/g00pMBnfrR>",
 "03": "Moderate earthquake - Molucca Sea on June 12, 2016 #tab <https://t.co/flstPxygxr>",
 "04": "Moderate earthquake - Bitung, Indonesia on June 12, 2016 #tab <https://t.co/lZsD0eEuge>"

Abbildung C.3: Fälschlich erkanntes Ereignis (Amoklauf) aufgrund von News. Ausschnitt Ausgabedatei mit Tweets, Keywords-Algorithmus.

"00": "We really are one shooting away from US troops being posted on every street corner. Hurry up and commission me, bro.",
 "01": "<https://t.co/ngIuF3M3> Child held hostage for third day after Alabama shooting: police: MIDLAN... <http://t.co/bm4p3Dcl> <http://t.co/nqIuF3M3>",
 "02": "Child held hostage for third day after Alabama shooting: police: MIDLAND CITY, Alabama (Reuters) - A standoff st... <http://t.co/o4igz0k>",
 "03": "Child held hostage for third day after Alabama shooting: police: MIDLAND CITY, Alabama (Reuters) - A standoff st... <http://t.co/9a0FQeX>",
 "04": "Chicago shooting victim Hadiya Pendleton starred in anti-gang video: Four years before she became the innocent v... <http://t.co/2R0Gj5IM>",
 "05": "Phoenix office shooting suspect possibly found dead <http://t.co/XrnK3hd>",
 "06": "Chicago shooting victim Hadiya Pendleton starred in anti-gang video: <http://t.co/YUgAGzi>",
 "07": "Chicago shooting victim Hadiya Pendleton starred in anti-gang video: Four years before she became the i... <http://t.co/NjGMBU3> (plz RT)",
 "08": "Now people just walking in shooting up district attorneys now my oh my",
 "09": "Child held hostage for third day after Alabama shooting: police: MIDLAND CITY, Alabama (Reuters) - A standoff st... <http://t.co/Frgj5kU8>"

Abbildung C.4: Fälschlich erkanntes Ereignis (Amoklauf) aufgrund von Ambiguität. Ausschnitt Ausgabedatei mit Tweets, Classification-Algorithmus.

"00": "This fucktard is shooting fire-crackers into our yard! I'm about to start catapulting rocks into his yard! #ComeAtMeBro !",
 "01": "Looking forward to @FromeNetball anyone interested I am running another shooting sess before main training x see you at 7pm x",
 "02": "Tonight I'll be shooting Bob Seger at the Veterans Memorial Arena for the first 2 songs, but we actually get to... <http://t.co/gmDSWt7Q>",
 "03": "Newt is shooting off his mouth without knowing specific facts. He is resorting to Democratic strategies in desperation. FAIL. @SHannitysHair",
 "04": "But meo #SoftPinkKisses u shud find out what shooting range n we shud go asapish!!!!!"

Abbildung C.5: Fälschlich erkanntes Ereignis (Erdbeben) aufgrund nicht abgedeckter Startzeitpunkte (Ongoing). Ausschnitt Ausgabedatei mit Tweets, Burstiness-Algorithmus.

"00": "Check this out! Nepal Is Devastated by Another Earthquake - On April 25th a devastating earthquake having a magnit... <http://t.co/LNi9V9bD7N>",
 "01": "Magnitude 6.8 quake hits off Japanese island of Honshu @timesofindia <http://t.co/cqJ3SWHx3>",
 "02": "Nepal earthquake: Rescue to resume after latest deadly tremor - BBC News: BBC NewsNepal earthquake: Rescue to ... <http://t.co/yqV64D4G6U>",
 "03": "With another earthquake in Nepal, Viber has made calling out of the country free <http://t.co/DkC3db1BxE> #tips",
 "04": "Nepal earthquake: Rescue to resume after latest deadly tremor - BBC News <http://t.co/xBjAMWfd6i>",
 "05": "Nepal monsoon season worries - Nepal's national nightmare continues after being hit by a second earthquake with a ... <http://t.co/N023G3pIdR>",
 "06": "Help support the victims of the Japanese Earthquake and Pacific Tsunami <http://t.co/811VNSR8> #hope4japan #pray4japan",
 "07": "Amount paid for a Picasso painting- \$1,793,300,000. Amount raised for 2 earthquakes in #Nepal-\$154,155,964.\nSurely you can do better, World!",
 "08": "Amazing #smile BBC NewsNepal earthquake: Rescue to resume after latest deadly tremorBBC NewsRescue work is se... <http://t.co/YKtG9wF1s2>",
 "09": "Second earthquake worries local Nepalese groups: Nepal was rocked again, just weeks after a devastating earthquake... <http://t.co/V1897Qdhj7>",
 "10": "prayer to suppress earthquakes <http://t.co/r13RYzTmg8>",
 "11": "Now the second earthquake. Lord father not even the strongest man can match your weakness, father we pray to you for things to be as youwish",
 "12": "Nepal earthquake: Rescue to resume after latest deadly tremor - BBC News: BBC NewsNepal earthquake: Rescue to ... <http://t.co/1M8a5FVtKk>",
 "13": "Did you guys feel that earthquake this morning!? I went back to sleep, of course!",
 "14": "6.8-magnitude earthquake hits Japan: The quake struck at 6:12 am (2112 GMT Tuesday) off the... <http://t.co/VpQVHwe5Jh>",
 "15": "MAGNITUDE 6.9 EARTHQUAKE HITS OFF MAIN JAPAN ISLAND OF HONSHU (USGS) <http://t.co/lgekm7EeF8>",
 "16": "Aftershock: #Kamloops Residents Urged to Donate to Earthquake Relief in #Nepal | VIDEO REPORT: <http://t.co/gEs2oUC0o2> via @TgostelowFCUC"