Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung

Nicolas With

Vergleich von Text Mining Algorithmen in Social Media in Bezug auf ein Epidemie-Frühwarnsystem

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Nicolas With

Vergleich von Text Mining Algorithmen in Social Media in Bezug auf ein Epidemie-Frühwarnsystem

Eingereicht am: 28. Februar 2013

Inhaltsverzeichnis

1 Einleitung **1**
1.1 Aufgabenstellung 1

2 Grundlagen **3**
2.1 Epidemie-Frühwarnsystem 3
2.2 Data-Mining vs. Text-Mining 4

3 Aufbau der Thesis **5**
3.1 Testumgebung 5
3.2 Text-Mining Algorithmen 6
3.3 Kriterien 8

4 Ausblick **9**
4.1 Chancen & Risiken 9

5 Zusammenfassung **10**

Abbildungsverzeichnis

2.1 Text-Mining Prozess 4

3.1 Architektur des M-Eco Systems 5
3.2 Die Architektur des geplanten Systems 6

1 Einleitung

In Zeiten, wo die Population auf der Erde rasant ansteigt und damit auch die Bevölkerungsdichte, nimmt auch die Gefahr einer landesweiten Epidemie zu. Wo mehr Menschen auf einem Fleck wohnen, steigt die Gefahr einer Ansteckung von Mensch zu Mensch und damit einer raschen Ausbreitung der Krankheit.

Durch bessere Technik im Bereich der Medizin ist es möglich bessere Impfstoffe und Gegenmittel herzustellen, um die rasche Krankheitsausbreitung schnell einzudämmen. Jedoch ist der einfachste Weg eine drohende Epidemie zu bekämpfen, jene so früh wie möglich zu erkennen, um somit schneller reagieren zu können, so dass bestenfalls die Impfstoffe anschlagen, bevor der Höhepunkt der Epidemie erreicht ist.

Solche Systeme, welche drohende Epidemien vorhersagen können, gibt es schon seit einiger Zeit. Diese Epidemie-Frühwarnsysteme arbeiten mit Daten, die aus vergangenen Ereignissen gesammelt wurden, um Aussagen über zukünftige mögliche Epidemien zu treffen. Da die Daten am Beginn des Internetzeitalters spärlich waren und die Ausbreitung neuer Krankheiten nicht verfolgt werden konnte, bezogen sich diese Systeme meist nur auf jährlich auftretende Influenza, wie die Grippe. Erst mit der voranschreitenden globalen Vernetzung stehen einem genug Daten zur Verfügung, um Erkenntnisse und Muster von seltener auftretenden Krankheiten zu sammeln. Foren, Blogs und soziale Netzwerke liefern dabei riesige Datenmengen, die es zu analysieren gilt.

Um die großen Datenmengen nach brauchbaren Informationen zu durchsuchen, werden effiziente Data-/ Text-Mining Algorithmen benötigt. Im Zuge des Mining-Prozesses werden die Daten aufbereitet, gefiltert, analysiert, interpretiert und schlussendlich präsentiert. Diese Daten können dann benutzt werden, um eine Warnung auszugeben, falls Daten gefunden wurden, die auf eine Epidemie schließen lassen.

1.1 Aufgabenstellung

Das Ziel der Masterthesis ist es, verschiedene Text-Mining Algorithmen, in Bezug auf ein Epidemie-Frühwarnsystem, zu untersuchen. Das Frühwarnsystem dient hier als Testumge-

bung, um die Wirksamkeit der Algorithmen zu überprüfen. Als Leitfaden zur Erstellung der Testumgebung dient das Projekt M-Eco der Universität Hannover.

Die Testumgebung wird soweit abgewandelt, dass sie generisch mit verschiedenen Algorithmen läuft, d.h. die Analyse- und Extraktionskomponente wird modular implementiert. Die Text-Mining Algorithmen, die behandelt werden, teilen sich in zwei verschiedene Bereiche auf, diese sind in Kapitel 3.2 näher erläutert. Die Kriterien, an denen die Algorithmen gemessen werden, sind unter Kapitel 3.3 aufgeführt.

Das erwartete Ergebnis ist eine Auflistung der unter den Gesichtspunkten der Kriterien untersuchten Text-Mining Algorithmen. Einerseits ein Vergleich in ihrem zugehörigen Bereich, andererseits ein Vergleich zwischen den Bereichen. Darüber hinaus soll eine lauffähige Version des Epidemie-Frühwarnsystems präsentiert werden, welche generisch mit den verschiedenen Algorithmen funktioniert und benutzt werden kann.

2 Grundlagen

2.1 Epidemie-Frühwarnsystem

Das Frühwarnsystem bildet den Rahmen der Masterthesis. Mit dem System wird versucht aus Daten, die aus sozialen Medien gewonnen werden, verlässliche Aussagen über Krankheitsverläufe und Krankheitsausbrüche zu machen. Dabei werden die gleichen Arbeitsschritte wie beim klassischen „Knowledge Discovery in Databases“ implementiert, nur übertragen auf soziale Medien und nicht auf Datenbanken oder im Data-Warehouse.

Die einzelnen Schritte des Prozesses sind wie folgt aufgebaut: Ester und Sander (2000)

- Daten erheben
- Daten aufbereiten
- Daten formatieren und speichern
- Daten analysieren
- Ergebnisse präsentieren

Im ersten Schritt werden Daten aus diversen Quellen gesammelt. Fachliche Quellen, wie medizinische Blogs und Medizinerforen, aber auch nicht-fachliche Quellen, wie Twitter, Facebook oder andere soziale Netzwerke werden durchsucht. Der folgende Aufbereitungsschritt nimmt einen großen Teil des Gesamtaufwands ein. Daten aus verschiedenen Quellen müssen integriert werden und auf eine gemeinsame Konvention gebracht werden. Inkonsistenzen, wie z.B. Schreibfehler, werden bereinigt und unwichtige Daten werden rausgefiltert. Denecke u. a. (2012)
Danach werden die Daten in ein geeignetes Format überführt und in einer Datenbank gespeichert. Die Daten werden dann analysiert. Dazu gibt es verschiedene Algorithmen, welche für unterschiedliche Aufgabenstellungen eingesetzt werden. Genaueres in Kapitel 3.2.
Zu guter Letzt werden die Ergebnisse dem User in einer geeigneten Form präsentiert, wobei dieser Schritt beim Text-Mining einen größeren Umfang einnimmt als beim klassischen Data-Mining.

2.2 Data-Mining vs. Text-Mining

Anders als Data-Mining ist Text-Mining kein eindeutig definierter Begriff. Im Rahmen dieser Arbeit wird Text-Mining definiert als Prozess, bei dem Algorithmen und Methoden aus dem Bereich des maschinellen Lernens und Statistik angewendet werden, mit dem Ziel nützliche Muster in unstrukturiertem oder semi-unstrukturiertem Text zu finden. Hotho u. a. (2005)

Obwohl andere Algorithmen aus anderen Fachbereichen im Text-Mining Verwendung finden, ist der Ablauf des Prozesses ähnlich der des „Knowledge Discovery in Databases“, weswegen in manchen Fällen von „Knowledge Discovery in Text“ gesprochen wird.
Abbildung 2.1 demonstriert den Ablauf, wie er üblicherweise im Text-Mining angewandt wird.

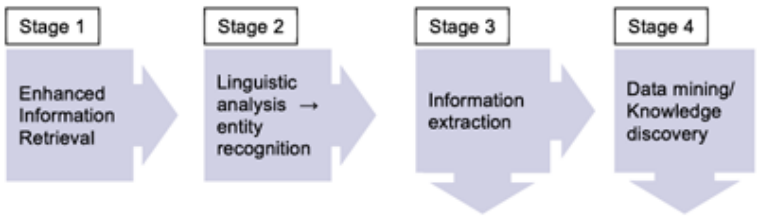

Abbildung 2.1: Text-Mining Prozess
McDonald und Kelly (2012)

Im ersten Schritt werden die Daten erhoben. Hier werden Techniken aus dem Bereich „Information Retrieval“ (IR) angewendet, wie sie z.B. auch in Suchmaschinen benutzt werden. Wichtig ist, dass es beim IR nur darum geht bestehende Informationen aufzudecken und nicht neue Strukturen zu finden, dies wird in den folgenden Schritten gemacht. Hotho u. a. (2005)
Nachdem passende Dokumente gefunden wurden, werden diese aufgearbeitet. Dabei kann eine Reihe von Pre-Processing Methoden angewendet werden, welche meistens aus dem Bereich „Natural language processing“ kommen. Darunter gehören Techniken wie filtern, Worte auf ihre Grundformen zurückführen oder *Keyword Selection algorithms*. Witte und Mülle (2006) Heyer u. a. (2008)
In dem darauf folgenden Schritt kann damit begonnen werden aus dem Text Informationen zu ziehen. Methoden, die hierbei zum Einsatz kommen, sind *Tokenization*, *sentence segmentation* und die Identifizierung von Personen, Orten, Organisationen und Zeiten. Diese Informationen werden in Datenbanken gespeichert. Hotho u. a. (2005)
Im letzten Schritt werden Data-Mining Algorithmen auf die in der Datenbank gespeicherten Informationen angewendet, um neue und nützliche Muster zu finden.

3 Aufbau der Thesis

3.1 Testumgebung

Als Blaupause für die Testumgebung dient das Projekt „Medical Ecosystem“ (M-Eco) der Uni Hannover unter der Leitung von Prof. Dr. Kerstin Denecke. M-Eco soll eine Plattform bieten, die an das bestehende MediSys System angeschlossen wird. MediSys durchforstet Nachrichten auf der ganzen Welt, um eine Übersicht und Warnungen über diverse Krankheiten zu liefern. Dieses System versucht M-Eco um soziale Medien und soziale Netzwerke zu erweitern, d.h. den Suchradius zu vergrößern. Die herausgefilterten Informationen werden in einer Datenbank gespeichert. Der Benutzer kann dann Suchanfragen erstellen, die z.B. Kranheiten, Beschwerden und Zeiträume beinhalten, worauf das System zur Suchanfrage passende Ereignisse liefert. Denecke u. a. (2012)

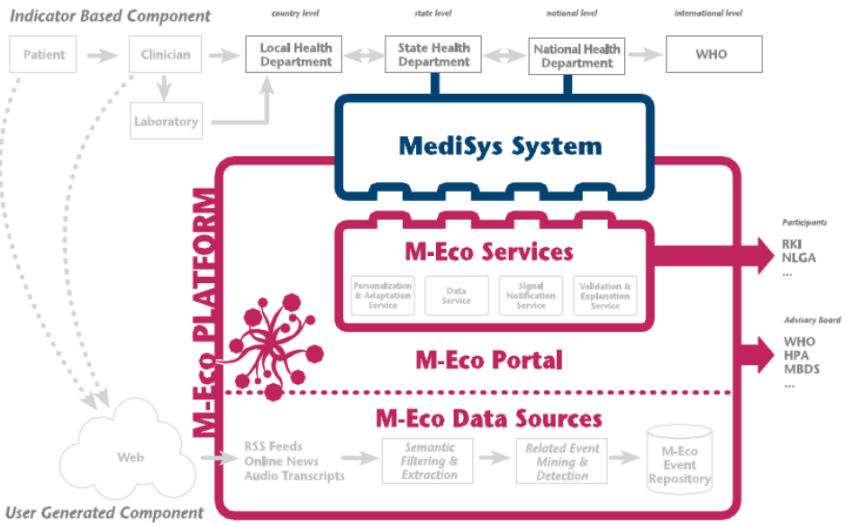

Abbildung 3.1: Architektur des M-Eco Systems
Denecke u. a. (2012)

Die Architektur des M-Eco-Systems wird in Abbildung 3.1 gezeigt. Es werden Daten aus Online-Nachrichten, Blogs und Twitter gefiltert. Zum Teil werden Daten über externe Tools gesammelt, wie z.B. „SAIL Media Mining System“ oder „MedWorm“, oder über korrespondierende APIs. Das Pre-Processing der Daten erfolgt durch „OpenCalais“ und „Minipar“, wobei ersteres die Dokumente nach relevanten Begriffen durchsucht und letzteres die Dokumente sprachlich analysiert und daraufhin strukturiert. Danach werden die Dokumente durch Lernalgorithmen in Cluster zusammengelegt und die Suchanfragen der User in Ergebnisse umgewandelt.

Die Module zur Informationsbeschaffung („Information Retrieval“) und zur Vorverarbeitung („Pre-Processing“) werden dem M-Eco-System entnommen. Der Analyseprozess wird ein modulares System, welches die einzelnen Algorithmen generisch benutzen kann, um somit einen bestmöglichen Vergleich anzustreben. Bei der Präsentation der Ergebnisse wird eine Ansicht der Events über Google Maps angestrebt. Im besten Fall wird für jedes Ereignis das gefunden wird, ein Punkt auf der Karte erzeugt, so dass die Ausbreitung oder der Verlauf einer Krankheit übersichtlich dargestellt werden kann. Der Aufbau des Systems wird in Abbildung 3.2 skizziert.

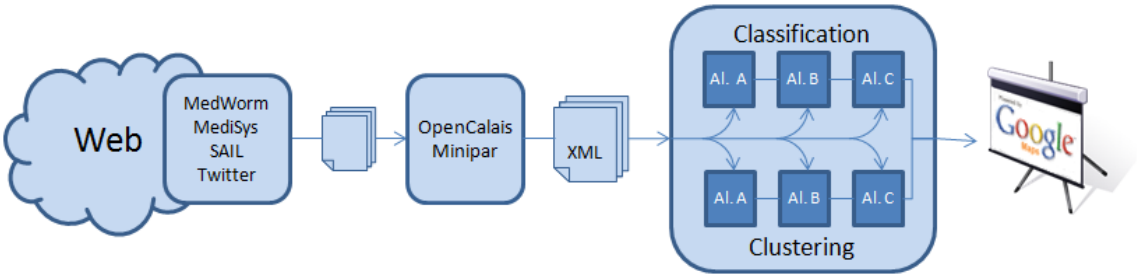

Abbildung 3.2: Die Architektur des geplanten Systems

3.2 Text-Mining Algorithmen

Wie im Kapitel 2 erläutert wurde, gibt es für die einzelnen Schritte eine Vielzahl von Algorithmen und Techniken, die angewendet werden können. Für diese Arbeit werden nur Techniken der beiden letzten Prozessschritte berücksichtigt. Die Dokumentenbeschaffung und Vorverarbeitung wird anhand von Tools bewerkstelligt, die in Kapitel 3.1 vorgestellt wurden.

Es wird dabei unterschieden zwischen *Klassifizierungsverfahren* und *Clusterverfahren*. Da es verschiedene Herangehensweisen gibt, wie die Idee, die hinter den Verfahren steckt, umgesetzt werden kann, wurden jeweils 3 Algorithmen ausgewählt.

- Classification
  - Naïve Bayes
  - Nearest Neighbor
  - Support Vector Machines
- Clustering
  - Bi-section k-means
  - Self organizing map
  - Fuzzy clustering

Die Algorithmen werden einerseits untereinander verglichen, andererseits werden die beiden unterschiedlichen Verfahren verglichen.
Classification beschreibt ein Verfahren, bei dem Dokumente in eine oder mehrere Klassen sortiert werden. Die Klassen sind dabei vordefiniert und werden mit Beispieldokumenten "trainiert". Um eine Zuordnung vorzunehmen wird ein neues Dokument nach Schlüsselwörtern durchsucht, die angeben, welche Klasse dem Dokument zugewiesen wird. Hotho u. a. (2005) Ein Beispiel für ein Klassifizierungsverfahren sind die Spamordner in E-Mail-Programmen, die anhand von definierten Regeln einer neuen Mail das Label „Spam" oder „Nicht-Spam" zuordnen.

Clusteringverfahren ordnen Dokumente in Cluster, wobei ähnliche Dokumente in gleichen Clustern liegen sollen und unähnliche Dokumente in verschiedenen Clustern. Die Qualität eines Clusterverfahrens kann daran gemessen werden, wie ähnlich die Dokumente in einem Cluster sind und wie sehr sich die Cluster voneinander unterscheiden. Witte und Mülle (2006) Clusterverfahren werden z.B. im Marketing eingesetzt, um Kunden in Gruppen aufteilen zu können, die dann mit gezielten Marketingstrategien angesprochen werden.

Der Unterschied zwischen den beiden Verfahren ist, dass die Klassifizierungsverfahren vordefinierte Klassen benutzen und somit eine statische Zuordnung vornehmen, während die Clusterverfahren eine dynamische Zuordnung der Objekte vornehmen. Im Kontext des maschinellen Lernens kann gesagt werden, dass Klassifizierungsverfahren mit *überwachtem Lernen* und Clusterverfahren mit *unüberwachtem Lernen* arbeiten. Wikipedia (2013)

3.3 Kriterien

Nachdem die Testumgebung und die ausgewählten Algorithmen implementiert wurden, müssen sie bewertet werden. Die Bewertung läuft anhand von Kriterien ab, die für Text- aber auch für Data-Mining Algorithmen wichtig sind. Zu den ausgesuchten Kriterien gehören:

- Ressourcenverbrauch
- Schnelligkeit
- Ergebnisqualität
- Ergebnisquantität
- Implementationsumfang

Der *Ressourcenverbrauch* gibt an, wie viel Speicher- und CPU-Last der Algorithmus verursacht. Dabei wird die Speicherlast in MB und die CPU-Last als durchschnittlicher Prozentanteil der gesamten CPU-Leistung angegeben. Zur Messung kann z.B. der windowseigene Task-Manager verwendet werden.

Mit der *Schnelligkeit* wird angegeben, wie schnell der Algorithmus Ergebnisse ausgibt, dabei ist die Ergebnisqualität irrelevant. Es zählt nur, wie viel Zeit benötigt wird, um eine bestimmte Menge an Ergebnissen geliefert zu bekommen. Gemessen wird in Millisekunden.

An der *Ergebnisqualität* lässt sich messen, wie hochwertig die Ergebnisse sind, die geliefert wurden. Dabei ist es unerheblich, wie viel Zeit in Anspruch genommen wurde, um die Ergebnisse zu liefern, solange sie sich in einem vorher definierten Grenzwert halten. Die Qualität der Ergebnisse zu bestimmen ist nicht so geradlinig, wie bei den anderen Kriterien, da es kein eindeutiges Maß für Qualität gibt. In diesem Fall sagt die Qualität aus, ob ein korrekter Schluss aus den Informationen gezogen wurde. Dies muss dann manuell nachgeprüft werden.

Die *Ergebnisquantität* orientiert sich an der Menge der zurückgelieferten Ergebnisse. Hierbei muss noch bestimmt werden, inwieweit ein Ergebnis als Ergebnis zählt, d.h. ab welcher Güte das Ergebnis als solches angenommen und gezählt wird.

Ein letztes wichtiges Kriterium ist der *Implementationsumfang*, der angibt in welcher Zeit und mit welchem Aufwand der Algorithmus in der Testumgebung implementiert wurde. Anders als die anderen Kriterien ist dies ein eher subjektives Kriterium, da es womöglich verschiedene Wege gibt einen Algorithmus zu implementieren, abgesehen davon, dass er in unterschiedlichen Programmiersprachen anders implementiert werden kann. Durch die subjektive Natur des Kriteriums gibt es kein eindeutiges Maß, an dem der Implementationsumfang gemessen werden kann, daher wird es eine textuelle Beschreibung geben.

4 Ausblick

4.1 Chancen & Risiken

Durch den Vergleich von Text-Mining Algorithmen im Umfeld der sozialen Medien kann herausgefunden werden, welche Algorithmen für die Extraktion von Wissen aus diesem Bereich geeignet sind. Dies ist nicht nur für das Szenario des Epidemie-Frühwarnsystems von Vorteil, sondern auch für etwaige andere Anwendungen, welche Informationen aus sozialen Medien brauchen.

Als schwierig wird sich herausstellen, genug Informationen zu bekommen, um nicht-saisonal auftretende Krankheiten erkennen zu können. Wenn der Radius auf Deutschland und somit auf deutsche Quellen beschränkt wird, wird es schwierig sein eine verlässliche Aussage über die Güte der Algorithmen zu treffen, d.h. die Analyse beschränkt sich entweder auf englischsprachige Texte oder es wird mehrsprachlich implementiert, was jedoch den Umfang der Arbeit vergrößern könnte.
Darüber hinaus könnte die Anzahl der zu überprüfenden Algorithmen zu Problemen führen, jedoch ist diese Zahl jederzeit anpassbar, da versucht wird die Analysekomponente des Systems modular zu gestalten.
Eine letzte Sache ist die, dass durch einen zu großen Fokus der Arbeit auf die Text-Mining Algorithmen das Frühwarnsystem, welches eigentlich der Mittelpunkt der Arbeit sein sollte, in den Hintergrund geraten und zu einem austauschbaren Szenario werden könnte. Es muss also versucht werden zwischen der Entwicklung des Systems und dem Vergleichen und Testen der verschiedenen Algorithmen eine Balance zu finden.

5 Zusammenfassung

In der Masterthesis geht es darum im Umfeld eines Epidemie-Frühwarnsystems, welches durch Text-Mining in sozialen Medien versucht Epidemien zu erkennen und voherzusagen, verschiedene Text-Mining Algorithmen anhand von Kriterien zu untersuchen und sie, falls möglich, zu vergleichen.
Die Testumgebung ist angelehnt an Arbeiten des Projekts M-Eco der Uni Hannover und wird generisch implementiert, um die verschiedenen Algorithmen einfach benutzen zu können.
Die Algorithmen können verschiedenen Techniken zugeordnet werden, die im Text-Mining Verwendung finden, namentlich der *Klassifizierung* und dem *Clustering*.
Die Kriterien, anhand denen die Algorithmen gemessen werden, sind *Ressourcenverbrauch*, *Schnelligkeit*, *Ergebnisqualität*, *Ergebnisquantität* und *Implementationsaufwand*.
Das Ergebnis sollte eine Auflistung und ein Vergleich der Algorithmen anhand der Kriterien sein und es sollte ein lauffähiges Epidemie-Frühwarnsystem herauskommen, welches selbstständig Quellen durchsucht, Information extrahiert und Muster erkennt, die es erlauben zukünftig weitflächigere Krankheitsausbrüche aufzuzeigen.

Literaturverzeichnis

[Denecke u. a. 2012] Denecke, Kerstin ; Dolog, Peter ; Smrz, Pavel: *Making Use of Social Media Data in Public Health*. WWW 2012 - European Projects Track. 2012

[Ester und Sander 2000] Ester, Dr. M. ; Sander, Dr. J.: *Knowledge Discovery in Databases*. Springer Verlag, 2000. – ISBN 3-540-67328-8

[Heyer u. a. 2008] Heyer, Gerhard ; Quasthoff, Uwe ; Wittig, Thomas: *Text Mining: Wissensrohstoff Text*. W3L-Verlag, 2008. – ISBN 3-937137-30-0

[Hotho u. a. 2005] Hotho, Andreas ; Nürnberger, Andreas ; Paass, Gerhard: *A Brief Survey of Text Mining*. LDV-Forum. 2005

[McDonald und Kelly 2012] McDonald, Dr. D. ; Kelly, Ursula: *The Value and Benefits of Text Mining*. JISC, 2012. – URL http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

[Wikipedia 2013] Wikipedia: *Statistical classification — Wikipedia, The Free Encyclopedia*. 2013. – URL http://en.wikipedia.org/w/index.php?title=Statistical_classification&oldid=540788197. – [Online; accessed 28-February-2013]

[Witte und Mülle 2006] Witte, Dr. R. ; Mülle, Jutta: Text Mining: Wissensgewinnung aus natürlichsprachigen Dokumenten / Universität Karlsruhe - Fakultät für Informatik. März 2006. – Forschungsbericht. Wintersemester 04/05