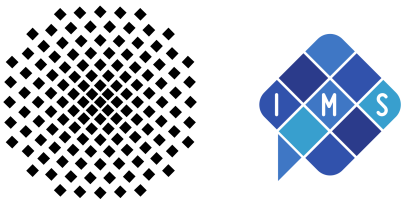
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Bachelorarbeit Nr. 328

Automatische Kategorisierung von Autoren in Bezug auf Arzneimittel in Twitter

Min Xu



Studiengang: Informatik

Prüfer: Prof. Dr. Jonas Kuhn
Betreuer: Dr. Roman Klinger
Beginn am: 09.11.2015
Beendet am: 09.05.2016

CR-Nummer: I.2.7, I.5.4, J.3

Kurzfassung

Mit der rasch wachsenden Popularität von Twitter werden auch immer mehr unterschiedliche Themen diskutiert. Dies lässt sich auch im Bezug auf die Wirkung von Arzneimitteln beobachten. Es ist daher sehr interessant herauszufinden, welche sozialen Gruppen dazu neigen, bestimmte Arzneimittel in Twitter zu diskutieren und welche Arzneimittel am meisten in Twitter diskutiert werden. Deshalb bietet es sich an, mit Verwendung der Technologie der Textklassifikation, die große Anzahl von Tweets zu kategorisieren. In dieser Arbeit wird das hauptsächlich mit dem Maximum Entropy Klassifikator realisiert, mit den sich die Autoren der Tweets erkennen lassen. Da das Maximum Entropy Modell eine Vielzahl der relevanten oder irrelevanten Kenntnis der Wahrscheinlichkeiten umfassend beobachten kann, erzielt der Maximum Entropy Klassifikator im Vergleich zum naiven Bayes-Klassifikator in dieser Arbeit ein besseres Ergebnis bei der Multi-Klassen-Klassifikation. Die Beeinflussung auf die Leistungen des Maximum Entropy Klassifikator unter der Verwendungen von verschiedenen Methoden, wie Information Gain & Mutual Information und LDA-Topic Model, zur Auswahl der Merkmale und unterschiedlicher Anzahl an Merkmalen wird verglichen und analysiert. Die Ergebnissen zeigen, dass die Methoden Information Gain & Mutual Information und LDA-Topic-Model gute praktische Ansätze sind, mit denen die Merkmale kurzer Texte erkannt werden können. Mit dem Maximum Entropy Klassifikator wird eine durchschnittliche Testgenauigkeit von 79.8% erreicht.

Abstract

With the rapidly growing popularity of Twitter there is also a growing amount of themes being discussed. This can also be observed relating to the effect of drugs. Therefore it is really interesting to figure out what social groups are tend to discuss drugs and what drugs are discussed the most in Twitter. To do so it makes sense to use the technology of text classification to categorize the huge amount of tweets. In this paper the detection of a tweet's autzhor is realized by the Maximum Entropy Classificator. The Maximum Entropy Modell is able to observe the variety of relevant an irrelevant acquirements of probability. It archieves better results compared to the Naive Bayes Classificator in multi-class-classification. The effect on the performance of the MaxEnt-classificator using different methods like Information Gain & Mutual Information and LDA-Topic Model for choosing characteristics and the use of different quantities of characteristics will be compared and analyzed. The results show that the methods Information Gain & Mutual Information and LDA-Topic-Model are good practical approaches for detecting characteristics of short texts. The test-precission of the Maximum Entropy Classificator reaches an average of 79,8%.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

List of Listings

1 Einführung

1.1 Motivation

Seit einiger Zeit legen die Menschen in vielen Ländern immer mehr Wert auf die eigene Gesundheit und eine gute medizinische Versorgung. Der damit steigende Verbrauch und Bedarf an Medikamenten, fördert die Forschung und Entwicklung von Medikamenten kontinuierlich. Damit wachsen die fachliche Literatur und Aufzeichnungen im Fachbereich der Medizin sehr schnell. Wir stehen vor solchen riesigen Datenmengen, dass händische Auswertung allein nicht ausreicht, wertvolle und umfangreiche Erkenntnisse zu gewinnen. Es besteht nun dringender Bedarf die Technologie der Text Mining für wissenschaftliche Texte und der Extraktion der Informationen über die Wirkung von Arzteinmitteln zu verbessern.

Dazu gibt es hier das Problem, dass viele wichtige Informationen einzelner Medikamenten nur von Ärzten, Pharmaunternehmen oder auch Behörden gesammelt und danach veröffentlicht werden. Heutzutage gibt es im Allgemeinen zwei Mechanismen, Nebenwirkungen eines Medikaments festzustellen [CCK14]. Der Erste sind klinische Studien die notwendig sind um ein neues Medikament auf den Markt bringen zu können. Den zweiten bilden Berichte von Agenturen, wie dem Zentrum für Krankheitskontrolle und Präventionen aus Krankenhäusern oder von Pharmaunternehmen usw. Die FDA (Food and Drug Administration) [1] verwaltet und überwacht zum Beispiel in den USA die Sicherheit der Medikamente auf dem Markt. Alle auf Nebenwirkungen bezogenen Informationen und Anwendungsfälle werden in einem System (Adverse Event Reporting System, AERS) der FDA gesammelt. Aufgrund der großen und schnell wachsenden Anzahl unterschiedlicher Arzneimittel selbst ist es mit diesen beiden Mechanismen allein schwer alle Information einzuholen. Dazu fehlt immer noch eine Methode, Informationen zu einem Medikaments direkt von Patienten zu erhalten.

Mit der Entwicklung des Internets und insbesondere der Popularität der Anwendung von Web 2.0 haben sich mannigfache Formen von User-Generated Content (UGC) wie Foren, Blogs, Twitter, Facebook usw. im Internet schnell verbreitet. Sie bereichern unser Netzwerk, bilden eine der wichtigen Wissensdatenbanken und spielen somit eine zunehmend wichtige Rolle. In Bezug auf die Gesundheitsversorgung sind in den letzten Jahren viele Webseiten, speziell für das Gesundheitswesen in soziale Medien erschienen, wie z.B. DailyStrength[2] und Health and Wellness Yahoo! Groups[3]. Allerdings sind solche Plattformen für die meisten Nutzer immer noch zu spezialisiert. Demgegenüber ist Twitter häufiger und beliebter verwendet.

Twitter[4] hatte 2013 schon über 200 Millionen registrierte Benutzer [5]. Die Benutzer verfassen

[1] www.fda.gov
[2] www.dailystrength.org
[3] https://groups.yahoo.com/neo/dir/1600060813
[4] www.twitter.com
[5] https://de.wikipedia.org/wiki/Twitter

eine Fülle von Informationen oder eigene Meinungen mit prägnanten Texten, die sie dann veröffentlichen und teilen. In Twitter werden Benutzererfahrungen, Kommentare und auch neueste Informationen von Medikamentenherstellern und neuestes Feedback von Nutzern oder von Verwandten und Freunden der Nutzer über die Wirksamkeit und Nebenwirkungen der Medikamente diskutiert und veröffentlicht. Dadurch können die Rückmeldungen über die Wirkung des Arzneimittels direkt von Patienten gewonnen werden.

In einer Umfrage ermittelten S.Fox und M.Duggan, dass im Jahr 2012 über 35% Erwachsenen in den USA im Internet Informationen über Arzneimittel, für die sie sich interessierten, bewertet haben [FD13]. Dies zeigt, dass das Internet ein immer wichtigeres Mittel wird, über das die Menschen die Informationen über Arzneimittel erhalten und kommunizieren können. In dieser Arbeit werden daher die Beiträge aus dem sozialen Medium Twitter (folgend „Tweets“gezeichnet) verwendet. Die Tweets enthalten die relevanten Informationen über Arzneimittel, die durch die Methode von Text-Mining kategorisiert werden.

Twitter wurde als Forschungsplattform dieser Arbeit gewählt, da Twitter über fast alle Eigenschaften von sozialen Medien verfügt:

1. Prägnante Inhalte. Größe wird auf 140 Zeichen begrenzt.
2. Massive Menge an Daten aus reicher Quelle.
3. Sehr schnelle Ausbreitungsgeschwindigkeit und Echtzeit-Update. Jeder kann zu jederzeit, von überall seine Meinung über die Twitter-App oder Twitter Webpage posten.
4. Vielfältiger Benutzer. Es gibt kaum Altersbegrenzung, keine Beschäftigungsbeschränkung und keine religiöse Beschränkung. Jeder kann ein oder sogar mehrere Konten registrieren.

Wegen der schnell wachsenden Anzahl an Tweets und der fehlenden Möglichkeit, diese geordnet anzuzeigen, fällt es dem Nutzer schwer, die benötigten Informationen aus den Tweets heraus zu filtern. Aus diesem Grund ist es notwendig die Tweets effektiv zu organisieren. Text Klassifizierung/Kategorisierung mithilfe maschinellem Lernen ist ein wichtiges Fundament für Text Mining. Sie kann in hohem Maße das Problem der ungeordneten Informationen lösen und hilft den Benutzern bei der Lokalisierung der richtigen erforderlichen Informationen. Automatische Kategorisierung ist eine leistungsfähige Methode zur Verarbeitung von massivem Informationszufluss, deren Entwicklung aus diesem Grund enorm an Bedeutung zugenommen hat.

1.2 Zielsetzung der Arbeit

Ziel dieser Arbeit ist es, verschiedene Klassifikatoren mit unterschiedlichen Algorithmen und Merkmalen zu entwickeln, mit denen Tweets automatisch in entsprechende Gruppen eingeordnet werden können. Die Gruppen, die vorher bestimmt werden, sind: Arzt, Behörden, Forscher, Journalist, Other, Patient, Pharmaunternehmen und Spam. Diese 8 Klassen können die typischen sozialen Gruppen relativ gut repräsentieren, die Meinungen über Arzneimittel posten würden. In der Pre-Studie Phase wird ein Java Projekt mit Twitter API implementiert,

um die Tweets von November 2008 bis November 2015 zu durchsuchen und die benötigte Tweets, die die 130 populärsten und meistverkauften (USA) Medikamente enthalten, im Corpus zu sammeln. In der Hauptstudie werden, basierend auf den Methoden des maschinellen Lernens, mehrere Klassifikatoren trainiert. Von den am häufigsten verwendeten Verfahren, wie Naive Bayes, Maximum Entropy, C4.5 Decision Tree wird eins ausgewählt, mit dem die ungeordneten Tweets effektiv und automatisch kategorisiert werden.

Mit dieser Arbeit wird es zum einen möglich sein aus unterschiedlichen sozialen Gruppen die zu bestimmen, welche die meisten Tweets zu einem bestimmten Medikament verfasst hat. Zusätzlich kann mit dem Ergebnis der Klassifikation bestimmt werden, wie viele wichtig nützliche Informationen über Arzneimittel man aus Twitter erhalten kann.

1.3 Gliederung der Arbeit

Diese Arbeit wird wie folgt gegliedert:

- **Kapitel 2 -** Grundlagen und verwandte Arbeiten: In diesem Kapitel werden zuerst die relevanten wissenschaftlichen Arbeiten dargestellt. Danach werden die Grundlagen bezüglich des Text Mining und die verwandte Technologie sowie die Liste der medizinischen Arzteinmitteln vorgestellt.
- **Kapitel 3 -** Corpus Erstellung: Die Sammlung, Bearbeitung, Normalisierung und Annotation des Corpus werden in diesem Kapitel dargestellt.
- **Kapitel 4 -** Trainieren des Klassifikators und entsprechende Implementierung: Hier werden die Methodik der Auswahl der Merkmale und der Aufbau der Klassifikatoren vorgestellt.
- **Kapitel 5 -** Corpus Analyse: Die manuell gekennzeichneten Tweets werden in diesem Kapitel evaluiert. Die Top-Arzneimitte jeder Klasse und die am meisten vorgekommenen Arzneimittel im Corpus werden gelistet. Die Top-Merkmale verschiedener Methode werden beispielsweise gezeigt.
- **Kapitel 6 -** Evaluation: Die Ergebnisse von verschiedenen Klassifikatoren mit verschiedenen Merkmalen sowie von allen Klassen vs. Pipeline werden hier gezeigt und verglichen.
- **Kapitel 7 -** Zusammenfassung: Im letzten Kapitel wird diese Arbeit zusammengefasst und ein Ausblick auf künftige Weiterarbeit oder Einsätze vorgestellt.

2 Grundlagen und Verwandte Arbeiten

2.1 Stand der wissenschaftlichen Forschung

2010 nutzten Leaman et al. [LWS+10] die Benutzerkommentare des gesundheitsbezogenen sozialen Mediums - DailyStrength[1] als Quelle ihrer Studie. Sie schafften die Anerkennung der Entitäten, indem sie die Ähnlichkeit zwischen den Inhalten der Bewertungen und den Namen der Nebenwirkungen mit der Methode „sliding window“ berechnet haben. Sie evaluierten und stellten die automatische Extraktion der Beziehungen zwischen Arzneimittel und ihrer Nebenwirkungen mithilfe des Corpus von DailyStrength auf. Das Ergebnis der Erkennung des gekennzeichneten Dateisets sind 78.3% Präzision und 69.9% für 73.9% F-Measure.

2012 haben Sampathkumar et al. 7916 Nachrichten/Beiträge aus dem medizinischen Forum *Meications.com* gesammelt [SLC12]. Von den 7916 haben sie 100 Nachrichten manuell gekennzeichnet und die restlichen wurden basierend auf Wörterbuch-Matching automatisch annotiert. Sie verwendeten Hidden-Markov-Modell, um die Nebenwirkungen der Medikamente von den automatisch annotierten Texten zu erkennen. Für 10-Cross-Validation erhielten sie die durchschnittliche F-measure von 86.4% und für die Testdatei von 73,2%.

Mit der Popularität von Microblog wird Twitter immer häufiger als Corpus für eine Forschung verwendet. A. Park und P. Paroubek [PP10] nutzten 300000 Tweets als Corpus für die Analyse des Sentiments und der Extraktion von Meinungen der Benutzer. Durch die linguistische Analyse und mit der Methode N-Gram haben sie die Merkmale ausgewählt. In ihrer Arbeit wurde ein naive Bayes-Klassifikator zur Erkennung der Emotionen in Twitter aufgebaut. Mit ihm lässt sich ermitteln, ob ein Text positive, negative oder neutrale Emotionen aufweist. Der, zu einem sehr guten Ergebnis kommende, Klassifikator wurde aus dem naiven Bayes-Klassifikator mit den Techniken N-gram und POS-tags entwickelt [PP10].

Die Kanouchi et al. [KKO+15] haben mit Hilfe von Twitter eine Gesundheitsüberwachung geschaffen, indem sie die Betreffe einer Krankheit oder eines Symptoms im japanischen Twitter abschätzten. Bei der Auswahl der Merkmale haben sie die Methoden Bag-of-Words und N-Gram sowie alle Feature verwendet. Ihr Ergebnis zeigte, dass die Identifizierung des Subjekts von der Krankheit oder dem Symptom unabhängig ist.

M. J. Paul und M. Dredze [AAD14] durchsuchten kürzlich gepostete Themen nach Unpässlichkeiten und sie entdeckten, dass über eineinhalb Millionen Tweets kleinere Krankheiten wie Erkältungen, Allergien, Übergewicht und Schlaflosigkeit usw. erwähnen. Damit haben sie uns die quantitativen Korrelationen mit öffentlichen Gesundheitsdaten und qualitativen

[1] http://www.dailystrength.org/

Bewertungen der Modellergebnisse gezeigt. Ihre Ergebnisse legen nahe, dass Twitter eine breite Anwendbarkeit für die öffentliche Gesundheitsforschung ermöglicht [PD11].

2.2 Technische Grundlagen

2.2.1 Twitter REST APIs

OpenAPI[2] ist eine übliche Anwendung basierend auf dem Modus SaaS (Software as a Service). Web-Dienstleister verkapseln ihre Services in einer Serie von APIs (Application Programming Interface) und machen deren Nutzung der Öffentlichkeit oder Drittanbietern zugänglich. Heutzutage ist OpenAPI als Basis zur Entwicklung von Internet-Online-Diensten eine sehr gute Wahl, die von immer mehr Unternehmen verwendet wird.

Twitter ist nicht nur im Bereich sozialer Medien, sonder auch unter Softwareentwicklern beliebt. Es bietet mittels OpenAPI viele Schnittstellen für Anwendungen, mit denen Entwickler Automatisierungen für die Extraktion von Twitter-Daten entwickeln können. Twitter Rest API[3] dagegen wird der Zugang zu den Daten eingeschränkt. Twitter Rest API hat einige Anwendungen, zur Vermeidung von böswilliger Nutzung, limitiert[4]. Zum Beispiel begrenzt „Rate limits“ die Anfragen auf 180 alle 15 Minuten.

Search API, ein kleiner Teil der Twitter Rest API, ermöglicht die Abfragen der Indizes der letzten oder populärsten Tweets und Ähnlichem. Aber im Gegensatz zur Smartphone-App oder zum Web-Client (*Twitter.com* Suchfunktion) erlaubt Twitter Search API das Nehmen von Samples aus Tweets, die in den letzten 7 Tagen veröffentlichten worden sind. Da die Search API die Aufmerksamkeit auf Relevanz und Unvollständigkeit richtet, um möglichst wenig Tweets und Benutzer aus Suchergebnissen zu verlieren, bietet Twitter zusätzlich noch die Streaming API an. Um die vergangene Tweets (älter als 7 Tage) zu erreichen, wurde in dieser Arbeit ein zusätzliches Projekt [5] verwendet. Es kombiniert simuliertes Scrollen im Web-Browser und die Nutzung der Search API um ältere Tweets zu erhalten.

2.2.2 Mallet

Mallet[6] ist ein, auf Java basierendes Paket, speziell für maschinelles Lernen. Mit Mallet lassen sich NLP (Natural Language Processing), Text Classification, Topic Modeling, Text Clustering und Feature Selection usw. ausführen. Das Mallet für diese Arbeit verwendet wird, hat zwei Gründe:

[2] https://en.wikipedia.org/wiki/Open_API
[3] https://dev.twitter.com/rest/public
[4] https://dev.twitter.com/rest/public/rate-limiting
[5] https://github.com/Jefferson-Henrique/GetOldTweets-java
[6] http://mallet.cs.umass.edu/

- Mallet enthält die meisten Algorithmen für die Kategorisierung. Es schließt Naive Bayes, C4.5 Decision Tree, Maximum Entropy, Boosting, HMM (Hidden Markov Model) und Bagging sowie AdaBoost usw. ein.

- Es gibt zwei Möglichkeiten zur Verwendung von Mallet — entweder die Implementierung mit Java Code oder mit Kommandozeile (Abbildung 2.1). Ein Java Code lässt sich auch in Kommandozeile schreiben.

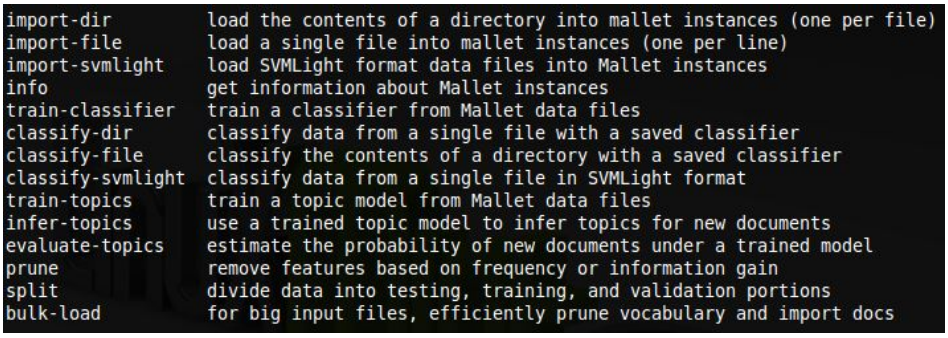


Abbildung 2.1: Beispiel: Mallet Commend Line

2.2.3 Liste der Arzneimittel

Für diese Arbeit wurden jeweils die Top 100 Arzneimittel der, in der *Pharmacy Times*[7] veröffentlichten, „Top 200 Products of 2011 by Total Prescriptions“ [8] und „Top 200 Products of 2011 by Total Dollars“[9], verwendet. Da viele Medikamente in den zwei Listen doppelt aufgetaucht sind, werden insgesamt 133 verschiedene Arzteinmittel erfasst.

Pharmacy Times ist ein von dem „the Best Healthcare Professional Media Brand award“ nominierte Monatszeitschrift[10]. Es stellt die in der täglichen Praxis benutzten klinischen Informationen für Pharmazeuten, Ärzten und medizinische Fachmenschen bereit. Jede Ausgabe enthält Artikel und Funktionen in Bezug auf Entwicklungstrends der Apotheke, Medikationsfehler, Nebenwirkung mit anderen Medikamenten, Patientenaufklärung, Pharmazie Technologie, Krankheitszustand Management, Patientenberatung, Produktneuheiten, die Arzneimittelgesetze und Spezial Apotheken.

In Anbetracht der unterschiedlichen sozialen Gruppen und Berufsfelder in dieser Arbeit, finden sich verschiedene Arzneimittelbezeichnungen für das gleiche Arzneimittel. Es ist daher hilfreich, die Synonyma der Arzneimittel herauszufinden.

[7] http://www.pharmacytimes.com/
[8] https://s3.amazonaws.com/pharmacytimes/d_media/_pdf/Top_200_Drugs_2011_Total_Rx.pdf
[9] https://s3.amazonaws.com/pharmacytimes/d_media/_pdf/Top_200_Drugs_2011_Total_Dollars.pdf
[10] https://en.wikipedia.org/wiki/Pharmacy_Times

DrugBank [11] ist eine Datenbank mit hilfreichen Ressourcen für Bioinformatik und Chemieinformatik. Sie bietet die umfassenden und detaillierten Informationen über die 7677 Arzneimittel [LKD+14]. Die Informationen in DrugBank werden immer aktualisiert, bis 2014 schon 4 mal [WKG+08] [KLJ+11]. Derzeit hat DrugBank über 6000 Datensätze der Arzneimittel und c.a zehntausende entsprechende Markennamen und Synonyma. Sie bietet zahlreiche Daten über Arzneimittel kostenlos herunterzuladen an.

Die Datenbank DrugBank der Universität Alterta bietet eine Quelle für Bioinformatik und Chemieinformatik an, die detaillierte Daten von Arzneimitteln (z. B. chemische, pharmakologische und pharmazeutische Daten) mit umfassenden Informationen zu Zielverbindungen (z. B. Sequenz, Struktur, Stoffwechselwege) verbindet. Die Datenbank enthält mehr als 6.800 Einträge. Zusätzlich sind mehr als 4.300 Proteinsequenzen mit diesen Einträgen verknüpft[12].

2.3 Grundlagen von Data Mining

Die grundsätzliche Aufgabe des Data Mining ist, versteckte interne nützliche Informationen oder wertvolle wissenschaftliche Kenntnisse von massiven ungeordneten Datenmengen herauszufiltern [TSK+06]. Hier werden ein paar Merkmale des Data Mining zusammengefasst[Nak13].

- Data Mining bezieht sich nicht nur auf statistische Theorie, sondern auf Datenbankmanagement, künstliche Intelligenz und Erkennung der Modells sowie Visualisierungstechniken und man muss bei der Verwendung eines Data-Mining-Tools keinen professionellen statistischen Hintergrund haben.
- Der Kern des Data Mining ist ein Algorithmus, welcher aber nicht übermäßig von einer strengen Schlussfolgerung abhängt, sondern von wesentlichen exploratorischen Ansätzen.
- Data Mining als Ergebnis von vielen Interdisziplinären hat im Vergleich zu Statistik mehr Praktikabilität, Exploration und Flexibilität. Data Mining ist ein induktives Verfahren.

In Bezug auf Data Mining sind in dieser Arbeit zwei Konzepte besonders wichtig und werden betont vorgestellt: Klassifikation (engl. Classification) und Vorhersagen (engl. Prediction). Klassifikation und Vorhersagen sind zwei Formen der Datenanalyse. Sie können verwendet werden, um die Menge wichtiger Features zu extrahieren und um zukünftige Trends eines Datenmodells vorherzusagen.

2.3.1 Klassifikationsalgorithmen und Klassifikator

Der Klassifikationsalgorithmus ist ein Verfahren, um das Problem der Kategorisierung zu lösen. Er spielt eine wichtige Rolle im Forschungsbereich von Data Mining, maschinelles

[11] http://www.drugbank.ca/
[12] https://de.wikipedia.org/wiki/DrugBank

Lernen und Mustererkennung. Durch die Analyse der Trainingsmenge der bekannten Klassen entdeckt der Klassifikationsalgorithmus die Regel der Kategorisierung. Damit prognostiziert die Klasse die neuen Daten.

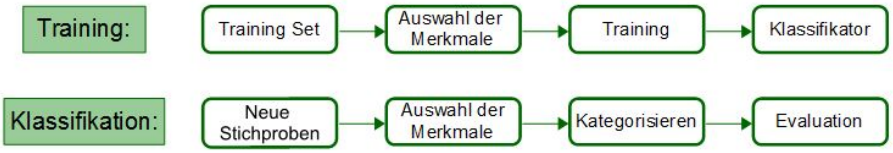


Abbildung 2.2: Training und Klassifikator

Naive Bayes Methode [MRS$^+$08]

Das Prinzip des Bayes-Klassifikators besagt, dass mit Hilfe der Priori-Wahrscheinlichkeit von einem Objekt seine Posteriori-Wahrscheinlichkeit nach der Formel von Bayes ausgerechnet wird. Die Posteriori-Wahrscheinlichkeit ist genau die Wahrscheinlichkeit, zu der das Objekt zu einer bestimmten Kategorie gehört. Die Kategorie mit der größten A-Posteriori-Wahrscheinlichkeit wird ausgewählt. Zur Zeit gibt es vier relativ bekannte und meist studierte Bayes-Klassifikatoren: Naive Bayes, TAN, BAN und GBN [DLY$^+$07]. Wir benutzen grundsätzlich nur Naive Bayes.

Die naive Bayes Methode basiert auf dem Satz von Bayes, nämlich:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}[\text{Wal}] \tag{2.1}$$

P(X | Y) ist die (bedingte) Wahrscheinlichkeit des Ereignisses X unter der Bedingung, dass Y eingetreten ist. P(Y | X) ist die (bedingte) Wahrscheinlichkeit des Ereignisses Y unter der Bedingung, dass XA eingetreten ist. P(X) ist die Wahrscheinlichkeit (Anfangswahrscheinlichkeit) für das Eintreten des Ereignisses X. P(Y) ist die Wahrscheinlichkeit (Anfangswahrscheinlichkeit) für das Eintreten des Ereignisses Y.

In der naiven Bayes Methode wird die Wahrscheinlichkeit, dass das Dokument d zur Kategorie c gehört, wie folgt berechnet [MRS$^+$08] :

$$P(c|d) \propto P(c) \prod_{lGkGn_d} P(t_k|c) \tag{2.2}$$

wobei $P(t_k|c)$ die bedingte Wahrscheinlichkeit ist, mit der t_k in den Dokumenten der Kategorie c zugewiesen wird. $P(c)$ ist die A-Priori-Wahrscheinlichkeit, mit der das Dokument der Kategorie c zugewiesen wird. Wenn anhand der Terme des Dokuments nicht klar zu erkennen ist, zu welcher Kategorie das Dokument gehören könnte, wird die Kategorie mit der größten A-Priori-Wahrscheinlichkeit ausgewählt. $< t_1, t_2, ..., t_n d >$ sind die Terme im Dokument d und n_d ist die gesamte Anzahl der Terme.

Das Ziel der Textklassifikation ist, die Kategorie mit der höchsten Wahrscheinlichkeit zu finden, der das Dokument zugeordnet werden soll. In der naiven Bayes-Klassifikation ist die Kategorie mit der höchsten Wahrscheinlichkeit genau das Ergebnis, das mit der MAP (maximum a posteriori) ist:

$$c_{map} = \arg\max_{c \in \mathcal{C}} \hat{P}(c|d) = \arg\max_{c \in \mathcal{C}} \hat{P}(c) \prod_{lGkGn_d} \hat{P}(t_k|c) \tag{2.3}$$

In der Formel wird ein geschätzter Wert $\hat{P}$, der durch Trainieren erhaltet wird, statt P verwendet, da der Realwert der Parameter unklar ist.

Die Formel 2.3 berechnet für alle k $(1 \leq k \leq n_d)$ die Produkte ihrer entsprechenden bedingten Wahrscheinlichkeiten. Diese Berechnungen verursachen eventuell einen Gleitkommaüberlauf. Deshalb wird das Logarithmieren eingeführt und die Formel 2.3 wird umgeschrieben:

$$c_map = \arg\max_{c \in \mathcal{C}} \hat{P}(c|d) = \arg\max_{c \in \mathcal{C}} [\log \hat{P}(c) + \prod_{lGkGn_d} \log \hat{P}(t_k|c)] \tag{2.4}$$

wobei der bedingt Parameter $\log \hat{P}(t_k|c)$ das Gewichte von t_k in der Kategorie c angibt und $\log \hat{P}(c)$ das Gewichte einer relativen Häufigkeit der Kategorie c. Im Vergleich zu den Kategorien mit niedriger Häufigkeit können die Kategorien mit hoher Häufigkeit eher richtig sein. Mit der Formel 2.4 wird die finale höchstwahrscheinliche Kategorie bestimmt.

Maximum Entropy [CT12]

Was ist Entropy? Der Begriff „Entropy“ wurde im Jahr 1864 von dem deutschen Physiker Rudolf Clausius vorgeschlagen. Er bezieht sich auf den zweiten Hauptsatz der Thermodynamik [13]. 1948 hat Informatiker C. Shannon diesen Begriff der Entropy erstmals in der Informationstheorie eingeführt. Die Entropy bedeutet die „Größe der Information“, das heißt, dass die Entropy ein Maß für die Menge der Informationen ist. Die Entropy einer Variable X beschreibt einen Wert der Unsicherheit von dieser Variable X. Gegeben ist die Definition der Entropy [14] wie folgt:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{2.5}$$

Je kleiner die Wahrscheinlichkeit jeder Zufallsvariablen X, desto größer ist die Entropy. Mit anderen Worten, je schwieriger eine Aufgabe ist, desto mehr Werte hat die entsprechende Lösung.

Was ist Maximum Entropy? Wenn wir die Wahrscheinlichkeitsverteilung eines Ereignisses vorhersagen müssen, sollten alle bereits bekannten Bedingungen erfüllt werden. Darüber hinaus dürfen keine subjektiven Annahmen an den unbekannten Umständen hinzugefügt werden. In diesem Fall ist die Wahrscheinlichkeitsverteilung am gleichmäßigsten und macht wenig Fehler bei der Vorhersage. Eine einfachere Erklärung des Prinzips von Maximum Entropy

[13] https://de.wikipedia.org/wiki/Entropie

[14] a. Wenn es für den Logarithmus zur Basis 2 ist, ist die Entropy in Bits; b. Wenn es für den Logarithmus zur Basis e ist, ist die Entropie in Nats.

(folgend „MaxEnt“ genannt) ist, wenn eine Unsicherheit auftaucht, sollen alle zugehörigen Möglichkeiten beibehalten werden, um das Risiko zu minimieren.

Maximum Entropy Modell ist die Theoriebasis der Maximum-Entropy-Klassifikator. Seine Grundidee ist, ein Modell für alle bekannten Faktoren zu erstellen, und alle unbekannten Faktoren auszuschließen. Das heißt, eine Wahrscheinlichkeitsverteilung herauszufinden, die alle bekannten Fakten erfüllt und nicht von irgendwelchen unbekannten Faktoren beeinflusst wird [BPP96]. Eine der bemerkenswertesten Eigenschaften des Maximum Entropy Modell ist, dass es keine bedingt unabhängigen Merkmale erfordert. Daher können wir die nützlichen Merkmale relativ einfach ohne Berücksichtigung der Wechselwirkungen zwischen ihnen in das System hinzufügen. Darüber hinaus kann man, im Vergleich zu den auf räumlichem Abstand basierenden Klassifikatoren wie SVM (Support Vector Machine), mit dem Maximum Entropy Modell leichter mehrere Klassen kategorisieren. Um mit dem SVM-Klassifikator eine Multi-Klassifikation realisieren zu können, muss sie dafür mit mehreren SVM-Klassifikatoren aufgebaut werden. Zum Beispiel werden $(K-1)$ SVM-Klassifikatoren für eine Klassifikation mit K Kategorien entwickelt [YL99]. Die Trainingseffizienz von Maximum Entropy ist im Vergleich zum SVM relativ gut, da die Trainingszeit und Klassifikationszeit von Maximum Entropy beide linear sind [Yan99].

Um die Merkmale im Maximum Entropy Modell formal auszudrücken, wird hierfür eine Merkmale-Funktion eingeführt [RJX+05]. Die Merkmale-Funktion ist eine Abbildung von ε nach $\{0,1\}$:

$$f_j(a,b) : \varepsilon \to \{0,1\} \tag{2.6}$$

In der Formel 2.6: $a \in A$, A repräsentiert eine Kategorie; $b \in B$, b ist ein Term aus Corpus B. Wenn $\tilde{p}(a,b)$ die empirische Wahrscheinlichkeitsverteilung ist, hat die Merkmal-Funktion einen Erwartungswert:

$$E_{\tilde{p}} f_j = \sum_{a,b} \tilde{p}(a,b) f_j(a,b). \tag{2.7}$$

Der Erwartungswert der Merkmal-Funktion bezüglich des Modells $p(a|b)$ ist:

$$E_p f_j = \sum_{a,b} \tilde{p}(b) p(a|b) f_j(a,b). \tag{2.8}$$

Während des Trainierens wird erzwungen:

$$E_{\tilde{p}} f_j = E_p f_j, \quad 1 \leq j \leq k \tag{2.9}$$

Nun verwandelt sich das Maximum Entropy Modell in eine Aufgabe, um eine Reihe von optimalen Lösungen zu finden, damit die Einschränkungen befriedigend sind. D.h ein $P^\star$ muss gesucht werden und folgendem entsprechen:

$$P = \{p | E_p f_j = E_{\tilde{p}} f_j, \quad j = \{1, ..., k\}\}, \qquad P^\star = argmax H(X) \tag{2.10}$$

davon ist $H(X)$ aus Formel 2.5.

Mutual Information [CT12]

Wie wir wissen, wenn $P(A, B) = P(A)P(B)$ gültig ist, dann sind A und B unabhängig. Die Beziehung zwischen P(A,B) und P(A) und P(B) kann durch Mutual Information (folgend „MI“ gezeichnet) gemessen werden. Bei Auswahl der Merkmale berechnen wir nämlich die MI zwischen Term t und Klasse c. MI beschreibt die Messung der Informationen, bei der eine Zufallsvariable eine andere Zufallsvariable enthält. Anders ausgedrückt, da andere Zufallsvariable die Information erhalten haben, wird die Unsicherheit der ursprünglichen Zufallsvariable reduziert. Um konkret zu sein, MI misst ob die Existenz des Terms t der Urteilsfähigkeit der Klasse c richtige Information bringen kann [MRS+08]. Die formale Definition wird wie folgend beschrieben:

$$I(U; C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)} \tag{2.11}$$

davon ist U eine binäre Zufallsvariable. Wenn das Dokument den Term t enthält, ist $e_t = 1$, sonst $e_t = 0$. C ist auch eine binäre Zufallsvariable, wenn das Dokument zur Klasse c gehört, ist $e_c = 1$, sonst $e_c = 0$. Wenn im Kontext die konkreten e und c nicht bestimmt sind, werden stattdessen U_t und C_c dargestellt.

Im Fall der Verwendung von MLE (maximum likelihood estimation) um die Parameter der Wahrscheinlichkeit abzuschätzen. Die Formel 2.11 ist äquivalent zur Formel 2.12.

$$I(U; C) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{10}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0.}N_{.0}} \tag{2.12}$$

Hier stellt jeder N_{xy} die entsprechende Anzahl der Dokumente im Fall von $x = e_t$ und $y = e_c$ dar. N_{10} gibt zum Beispiel die Anzahl der Dokumenten an, die den Term t ($e_t = 1$) enthalten aber nicht zur Klasse c gehören. $N_{1.} = N_{10} + N_{11}$ zeigt die Anzahl aller Dokumente, die den Term t enthalten. $N = N_{00} + N_{01} + N_{10} + N_{11}$ stellt die Anzahl aller Dokumente in allen Klassen dar. Wenn wir k Terme aus bestimmten Klassen auswählen, berechnen wir die $I = (U_t, C_c)$ für jeden Term und wählen die größten k Terme.

Information Gain [TSK+06]

In Information Gain (folgend „InfoGain“ gezeichnet) betrachtet man zwei Fälle, wobei ein Merkmal auftritt oder nicht auftritt. Es ist relativ vollständig und damit wird eine gute Wirkung erzielt. Vom Ganzen ausgehend: InfoGain untersucht die Kontribution eines Merkmales für das gesamte System, anstatt spezifisch für eine bestimmte Klasse. Auf Basis von Entropie in Kapitel 2.3.1 kann die Entropie für das Klassifizierungssystem wie folgend dargestellt

werden:

$$H(C) = -\sum_{i-1}^{n} P(C_i) \cdot \log_2 P(C_i) \tag{2.13}$$

In 2.13 bedeutet C eine Klasse, es ist eine Variable, d.h. die möglichen Werte für C sind $C_1, C_2, ..., C_n$. Die Wahrscheinlichkeiten, die in jeder Klasse auftreten sind: $P(C_1), P(C_2), ..., P(C_n)$. n ist die Gesamtzahl der Klassen. Der InfoGain eines Merkmals T für das gesamte System kann als die Differenz zwischen der Entropie ursprünglichen Systems und bedingter Entropie von T beschrieben:

$$\begin{aligned} IG(T) &= H(C) - H(C|T) \\ &= -\sum_{i-1}^{n} P(C_i) \cdot \log_2 P(C_i) + P(t) \sum_{i=1}^{n} P(C_i|t) \log_2 P(C_i|t) + P(\bar{t}) \sum_{i=1}^{n} P(C_i|\bar{t}) \log_2 P(C_i|\bar{t}) \end{aligned} \tag{2.14}$$

In Formel 2.14 ist T ein Merkmal und t bedeutet, dass das Merkmal T im Dokumentsystem auftritt und $\bar{t}$ heißt einen Gegenfall. Dementsprechend bedeuten $P(t)$ und $P(\bar{t})$ jeweils deren Wahrscheinlichkeit.

Latent Dirichlet Allocation (LDA)

Die Texte in dieser Arbeit sind sehr kurz, nämlich jeweils nicht länger als 140 Zeichen. Im Vergleich zu langen Texten haben kurze Texte eine Hauptschwierigkeit — die wichtigsten Merkmale kurzer Texte sind sehr spärlich und stark kontextsensitiv [LHPD12].

LDA bezeichnet ein unüberwachtes maschinelles Lernen um die versteckten Themen/Informationen in einem Corpus automatisch zu identifizieren [Ble][BNJ03]. Nachdem die Themen (-Verteilung) herausgefunden wurden, lassen sich damit die Clustern oder Textklassifikation weiter durchführen. LDA verwendet das Bag-of-words Modell, wobei jedes Dokument als ein Worthäufigkeitsvektor hingestellt wird Dadurch transformiert es die Textinformationen in digitale Informationen, um den Aufbau einer Datenmodellierung zu erleichtern. Jedes Dokument stellt eine Wahrscheinlichkeitsverteilung von einigen Themen dar, und jedes Thema umfasst die Wahrscheinlichkeitsverteilung von vielen Wörter. In Bag-of-words Modell berücksichtigen wir keine sequentielle Beziehung zwischen Wörtern und jedes Dokument könnte mehrere Themen enthalten. Für jedes Dokument Korpus definiert LDA ein generatives Verfahren:

1. Für jedes Dokument extrahiert es ein Thema aus der Themenverteilung.

2. Auswahl eines Wortes aus der Wortverteilung, die dem oben genannten Thema entspricht.

3. Wiederholung des 1. und 2. Punktes, anschließend Iteration aller Wörter im Dokument.

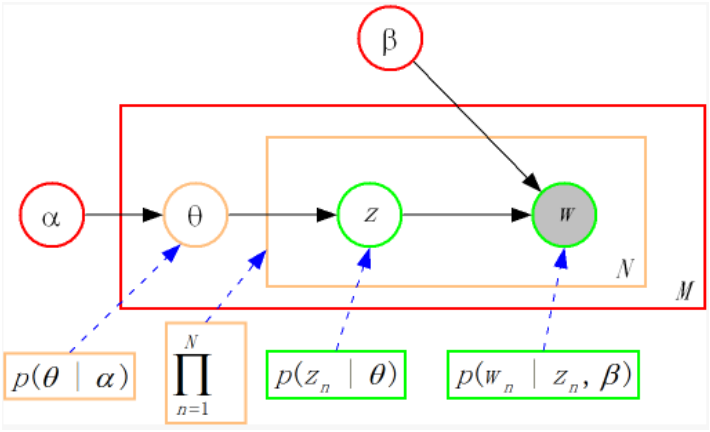


Abbildung 2.3: LDA Verfahren

In Abbildung 2.15 wird das Verfahren mit drei Farben gekennzeichnet:

1. Für Corpus (rot): α und β repräsentieren die Parameter der Corpus-Ebene, die sind jedem Dokument äquivalent.
2. Für Dokument (orange): θ ist eine Variable von Dokumenten. Jedes Dokument entspricht einem θ , und die von einem Dokument erzeugte Wahrscheinlichkeit für Thema Z ist jeweils anders.
3. Für Wort (grün): Z und W sind Variablen der Wortebene. Z wird von θ hervorgebracht, W wird von Z und β produziert. Ein Wort W entspricht einem Thema Z.

Davon bedeuten α und β :

- α : Verteilung $p(\theta)$ benötigt einen Vektorparameter, nämlich den Parameter für Dirichlet-Verteilung.
- β : Jedes Thema entspricht der Verteilungsmatrix $p(w|z)$ der Wortwahrscheinlichkeit.

Multivariate Wahrscheinlichkeit von LDA wird beschrieben wie folgend:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2.15}$$

2.3.2 Evaluation von Klassifikationen

Die Evaluation eines Klassifikators ist genauso wichtig wie der Klassifikator selbst. *Niemand ist perfekt, Jeder macht Fehler* — Kein Klassifikator kann in alle Fällen hundertprozentig korrekt funktionieren. In jedem Algorithmus können einige Fehler bei Klassifikationen auftreten und im Fall von „Big Data“ können die Daten oder Kategorien selber zweideutig sein. Deshalb ist es sehr wichtig, vor der praktischen Anwendung des Klassifikators, zuerst zu evaluieren

[MRS+08]. Es gibt einige Normen für die Beurteilung: Genauigkeit (engl. Accuracy), Präzision (engl. Precision), Rückruf (engl. Recall) und F1-Measure [BYRN+99].

Konfusionsmatrix (engl. Confusion Matrix)

Konfusionsmatrix ist eine spezifische Matrix um die Leistung des Algorithmus zu visualisieren und wird normalerweise unter überwachtem Lernen (engl. Supervised learning) verwendet. In der Matrix stellen jede Zeile/Spalte jeweils die tatsächliche/vorhergesagte Klasse dar. Sie zeigt deutlich, ob mehr als eine Kategorie miteinander vertauscht werden. Das heißt, eine Klasse wird in eine anderer Klasse vorhergesagt [Pow11].

		Vorhersagen	
		positiv	negativ
Tatsächliche Klassen	positiv	TP	FN
	negativ	FP	TN

Tabelle 2.1: Konfusionsmatrix für 2 Klassen

In Tabelle 2.1 eines binären Klassifikators sind die vier Begriffe wichtig:

Richtig positiv (TP) : Die tatsächliche Klasse ist positiv und der Test hat sie richtig angezeigt.

Falsch negativ (FN) : Die tatsächliche Klasse ist positiv aber der Test hat sie fälschlicherweise als negativ eingestuft.

Falsch positiv (FP) : Die tatsächliche Klasse ist negativ aber der Test hat sie fälschlicherweise als positiv eingestuft.

Richtig negativ (TN) : Die tatsächliche Klasse ist negativ und der Test hat sie negativ angezeigt.

Wenn es mehrere Klassen im System gibt (z.B wie Tabelle 2.2 gezeigt), muss man für jede Klasse die TP, FN, FP und TN getrennt rechnen. Beispielweise für Klasse X, sind ihre FN, FP und TN jeweils (XY + XZ), (YX + ZX), (TP von Klasse Y + TP von Klasse Z + YZ + ZY).

		Vorhersagen		
		Klasse X	Klasse Y	Klasse Z
Tatsächliche Klassen	Klasse X	TP	XZ	XZ
	Klasse Y	YX	TP	YZ
	Klasse Z	ZX	ZY	TP

Tabelle 2.2: Konfusionsmatrix für 3 Klassen

F-Measure, Presision, Recall [15]

Precision P: Anteil der tatsächlich positiven Klassen an allen als positiv vorhergesagten Klassen.

$$P(Precision) = \frac{TP}{TP + FP} \tag{2.16}$$

Recall R: Anteil der als positiv vorhergesagten Klassen an allen tatsächlich positiven Klassen.

$$R(Recall) = \frac{TP}{TP + FN} \tag{2.17}$$

Accuracy: Anteil der korrekten Entscheidungen an allen getroffenen Entscheidungen.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.18}$$

Accuracy wird sehr häufig für Maschinenlernprobleme verwendet. Sie ist nicht gut genug, um als eine Norm für Klassifikation-Evaluation oder eines Klassifikationsalgorithmuses zu gelten, da in den meisten Fällen die Daten im System extrem ungleichgewichtig sind. Zum Beispiel hat Google 100 Webseiten von DrugBank gecrawlt und es gibt insgesamt 10.000.000 Indexseite im System. Davon ziehen wir zufällig eine Seite um zu kategorisieren. Ist sie eine Webseite DrugBank oder nicht? Nehmen wir Accuracy um unsere Normen zu beurteilen, werden dann alle Webseiten als „Nicht von DrugBank“ markiert. In diesem Fall kann Accuracy derartig hoch bis $99,999\%(9.999.900/10.100.000)$ gehen.

In der Praxis werden Präzision und Recall beide mit höheren Werten erwartet, z.B wenn die FN und FP gleich 0 sind. Manchmal sind Präzision und Recall jedoch gegenseitig auszuschließen, z.B in dem Fall von 50 positive Proben und 50 negativen Proben, wo alle Ergebnisse als positiv vorhergesagt werden, dann sind Präzision und Recall jeweils 1 und $0,5$. Daher besteht hier ein Bedarf für einen Kompromiss zwischen Präzision und Recall — F1-Measure ist ein harmonisches Mittel aus P und R.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.19}$$

Kreuzvalidierungsverfahren, eg. Cross-Validation [K+95]

Kreuzvalidationen haben verschiedene Formen. In dieser Arbeit wird K-fache Keuzvalidation [Moo01] verwenden. In K-fache Kreuzvalidierungsverfahren wird ein Training Set in K Stücke geteilt. Eines davon wird als Verifikationsdaten beibehalten, die andere $K - 1$ Stücke sind zum trainieren. Jedes Stück wird einzeln verifiziert und k Wiederholungen berechnen

[15] Inhalt aus Folien der Vorlesung „Text Mining and Information Retrieval“ von Dr. rer. nat. Roman Klinger, IMS, University of Stuttgart

den Durchschnittswert. Jedes Stück wird zufällig generiert und zum Trainieren und auch zur Verifikation verwendet.

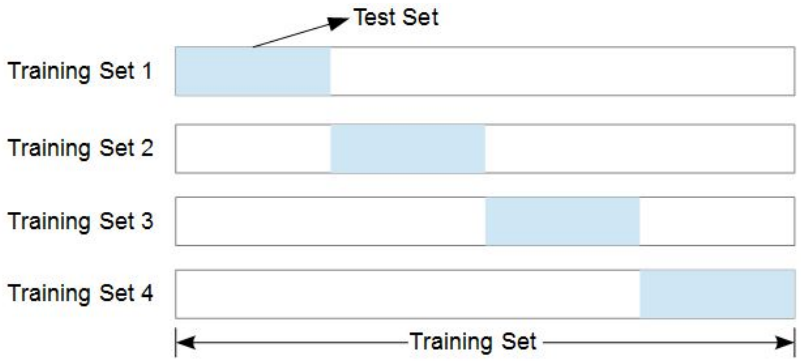


Abbildung 2.4: 4-fach Kreuzvalidierungsverfahren [GO06]

3 Corpus Erstellung

In diesem Kapitel wird das Corpus, das in den nächsten Kapiteln als Basis verwendet wird, dargestellt. Die zufällig ausgewählten Tweets werden nach entsprechender Methodik jeweils den vorgegebenen Klassen zugeordnet.

3.1 Sammlung des Corpus

Twitter API unterstützt mehrere Sprachen. Da wir die Region dieser Studie innerhalb aller englischsprachigen Länder bestimmt haben, wurden deshalb nur englische Tweets gecrawlt. Um die Daten der Studie möglichst allgemein und objektiv zu halten und so weit wie möglich nicht von besonderen Zeiten und Events beeinflusst, wurden die 205000 Tweets von April 2008 bis November 2015 gesammelt und davon 3000 Tweets zufällig ausgewählt.

Tweets	Arzteinmittel
@Mirikun: I've been on a ton of meds but now after a few months on **venlafaxine** I am absolutely astounded by the effects it has had	Venlafaxine
@mickeydoll94: PRILOSEC GENERIC OTC **Omeprazole** 20mg (120 pills) ***LOWEST PRICE***: http://bit.ly/si9yQu	Omeprazole
@xanaxhydrocodon: Mom had to bring me a **hydrocodone** to get rid of the pains: Mom had to bring me a... Buy **Hydrocodone** Here http://bit.ly/ot8IOx	Hydrocodone
@lvrampage: I miss **Detrol LA** commercials	Detrol LA
@Statins_bio: Effect of Evolocumab or **Ezetimibe** Added to Moderate or HighIntensity Statin Therapy on LDLC Lowering in Patients W... http://ow.ly/2GHfNs	Zetia

Tabelle 3.1: Einige Beispiele der originalen Tweets, die die Arzneimitteln enthalten.

3.2 Bearbeitung und Normalisierung des Corpus

Twitter hat ein 140-Charakter Limit für jeden Tweet. Diese reichen dem Nutzer oft nicht aus, um Information, die sie z.B. aus einer externen Quelle gewonnen haben, in einem Tweet auszuformulieren. Aus diesem Grund nutzen viele Nutzer die Möglichkeit, eine URL am Ende eines Tweets mit anzugeben. Über die URL hat der Leser die Möglichkeit sich auf z.B. die Informationsquelle verweisen zu lassen. Genauso erlaubt Twitter den Benutzern aber auch eigene Fotos oder Fotos von anderen Medien in Twiter zu teilen, weshalb viele URLs in

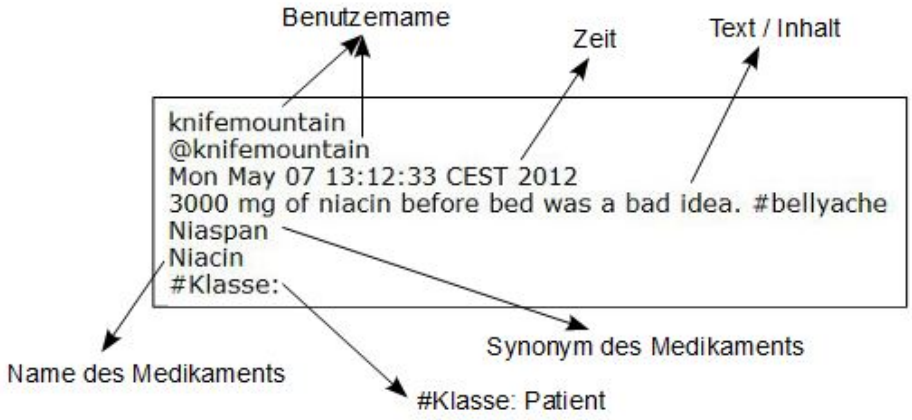


Abbildung 3.1: Konkretes Beispiel der Darstellung eines Tweets in dieser Arbeit

Tweets auftauchen. Es gilt nun die URLs, die relevante Informationen enthalten, von denen, die auf irrelevante Inhalte verweisen, zu separieren. Um die nachfolgende Vektorisierung mit unigrams, bigrams und trigrams zu erleichtern, werden die Tweets so bearbeitet, dass die Inhalte nicht geändert werden.

1. Übliche URLs wie *http://goo.gl/fb/umLv7* werden von **URLXXX** ersetzt; URLs von Fotos wie *pic.twitter.com/Y9ZZBKwloW* und aus Instagram wie *http://instagram.com/p/rf2CrCitnt/* werden jeweils von **URLPIC** und **URLINSTA** ersetzt.

2. Benutzernamen z.B „ @Alex“ (mit dem Symbol @ plus einen Benutzername) werden durch **@username** generalisiert. Wenn der Benutzername zufällig ein Name eines Medikaments ist (z.B @trazodone_), wird er mit **@Drugname** ersetzt.

Abbildung 3.2: Beispiele der Tweets nach Bearbeitung

3.3 Vorgegebene Klassen

Arzt: Die Ärzte, Krankenschwestern, fachliche Mitarbeiter in Krankenhäusern, Pharmazeuten, Apothekern sowie Unikliniken usw. werden in dieser Klasse eingesetzt.

Behörden: Nationale oder internationale Institutionen und Abteilungen bezüglich Arzteinmitteln wie z.B FDA (eg. U.S. Food and Drug Administration[1]) gehören zu dieser Klasse.

Forscher: Dieser Klasse enthält die medizinische oder pharmazeutische Professoren oder Studenten, die Forscher von medizinischer oder pharmazeutischer Instituten, Biologen, Chemiker oder Pharmakologen usw.

Journalist: Diese Klasse umfasst Nachrichtenmedien, Zeitungen und Zeitschriften sowie Gesundheitsblogs usw.

Patient: Patienten, Freunde oder Verwandte der Patienten werden dieser Klasse zugeordnet.

Pharma: Enthält die Pharmaunternehmen/Pharmafirmen.

Spam: Werbungen und Microblog-Marketing usw.

Other: Diese Klasse enthält die Tweets, deren Inhalte mit der Wirkungen von Arzteinmitteln nichts zu tun haben, oder die Tweets, die zu keiner oben genannten Klassen gehören.

Klassen	Tarining Set	Test Set	Sum
Arzt	438	47	480
Behörden	154	18	172
Forscher	382	41	421
Journalist	437	47	482
Patient	460	50	510
Pharmaunternehmen	226	25	251
Spam	408	45	452
Other	129	14	142
Sum	2634	287	2920

Tabelle 3.2: Verteilung des annotierten Corpus

3.4 Annotation

Bei dem Verfahren der manuellen Annotation wurden die Tweets anhand der Inhalte und dem Profil des Nutzers in Twitter (Verfasser des Tweets) der richtigen Klasse zuordnet. D.h der Benutzername eines Tweets wurde in das Suchfeld eingegeben und auf Twitter gesucht. In dem Profil des Nutzers wird in den meisten Fällen seine Beschäftigung beschrieben. Wenn das Profil leer ist oder keine nützlichen Informationen enthält, muss man andere Tweets des Nutzers untersuchen. Hier werden einige Beispiele gezeigt.

[1] http://www.fda.gov/

Tweet Anecdotal evidence that VEGF involed in lympedema development-breast cancer patients on bevacizumab report lymphedema improves.
Benutzername @K_BasenEngquist
Profil Director of Center for Energy Balance in Cancer Prevention and Survivorship at MD Anderson. I tweet on diet, exercise, healthy lifestyle, & cancer survivorship.
Klasse Arzt
Bemerkung Der Autor dieses Tweet ist (nach dem Profil) Direktor des Zentrums für Energiebilanz in der Krebsprävention

---

Tweet Intranasal midazolam vs rectal diazepam for paed seizure: terminates seizure faster, fewer adverse effects, parents prefer # FOAMped # PEM2013
Benutzername @_NMay
Profil Emergency Medicine/Paed EM doc, medical education enthusiast & # FOAMaoke queen. Retrieval reg @SydneyHEMS. Opinions mine, tweets for HCPs, here for the # FOAMed
Klasse Arzt
Bemerkung Der Autor dieses Tweet beschäftigt sich (nach dem Profil) mit der Notfallmedizin.

Tweet Testosterone Products: Drug Safety Communication - FDA Investigating Risk of Cardiovascular Events URLXXX # FDA
Benutzername @FDAMedWatch
Profil Clinically important safety information on human medical products from FDA. Comments: MedWatchComments@fda.hhs.gov Privacy: http://www.fda.gov/privacy
Klasse Behörden
Bemerkung Der Account dieses Tweet postet Informationen der FDA (Food and Drug Administration).

Tweet Phase IIIB Trial of Three Bortezomib-Based # Myeloma Regimens URLXXX # JCO # MMSM
Benutzername @ASCO
Profil ASCO is the American Society of Clinical Oncology. Making a world of difference in cancer care.
Klasse Behörden
Bemerkung Der Account dieses Tweet repräsentiert American Society of Clinical Oncology.

Tweet A Phase I Study of Capecitabine Combined with CPT-11 in Metastatic Breast Cancer Pretreated with Anthracycline... URLXXX
Benutzername @GoToPER
Profil Physicians' Education Resource (PER) specializes in oncology and hematology CME-certified activities. We advance cancer care through professional education.
Klasse Forscher
Bemerkung Der Account dieses Tweet spezialisiert auf Ontologie und Hämatologie CME-zertifizierten Aktivitäten.

Tweet Evotaz (atazanavir/cobicistat) gains positive opinion from CHMP for tx of HIV-1 pts without atazanavir resistance mutations # HIV # pharma
Benutzername @datamonitor_MH
Profil Senior Analyst at Datamonitor Healthcare. Views expressed are my own.
Klasse Forscher
Bemerkung Der Autor dieses Tweet ist (nach dem Profil) einen Senior Analyst bei Datamonitor Healthcare.

Tweet BioPortfolioNews 4811 HiTech receives tentative FDA approval for levofloxacin oral solution URLXXX BioPortfolioNews
Benutzername @BioPressRelease
Profil Track the latest biotechnology news, research, clinical trials, companies and reports. Continuously updated from 500+ news, research publications. BioPortfolio
Klasse Journalist
Bemerkung Der Account dieses Tweet postet die Nachrichten und Informationen der Medizin.

Tweet Zoloft is the most likely culprit, combined with clonazepam, together their similar oral side-effects combine URLXXX
Benutzername @eHealthMe
Profil eHealthMe provides health professionals and patients personalized tools to study millions of drug reports from FDA and community in real time.
Klasse Journalist
Bemerkung Der Account dieses Tweet postet die Nachrichten und Informationen der Medizin.

Tweet I used Claritin for three days as a child I mother was furious she wasted her money on that mess lol
Benutzername @_TOXICKisses_
Profil Cee
Klasse Patient
Bemerkung Aus Tweet ersichtlicher Klasse.

Tweet This hydrocodone is doin work on my mental faculties. Better that than intense abdominal pain though
Benutzername @feetypjz
Profil leer
Klasse Patient
Bemerkung Aus Tweet ersichtlicher Klasse.

Tweet Mylan Launches Generic Version of Plavix® Tablets URLXXX
Benutzername @MylanNews
Profil Official page for Mylan, where we work around the clock and around the globe to help provide 7 billion people access to high quality medicine
Klasse Pharma
Bemerkung Der Account repräsentiert das Unternehmen Mylan.

Tweet Our new # type2diabetes combo treatment w/ @username is available in U.S. pharmacies. Learn more: URLXXX # T2D
Benutzername @LillyDiabetes
Profil Welcome to the official Twitter profile of Lilly Diabetes. We're all diabetes, all the time!
Klasse Pharma
Bemerkung Der Account repräsentiert das Unternehmen Lilly.

Tweet hair removal treatment: 5 mg naltrexone , long term effects of hydrocodone use buy online cheap naltre Cymbala sonyericsson
Benutzername @CSENO
Profil Información desde Foro CSE. Todo para tu SonyEricsson.
Klasse Spam
Bemerkung Der Autor dieses Tweet postet Werbungen.

Tweet Mail order cheap Microzide without a prescription USA and worldwide URLXXX
Benutzername @Eslam2050
Profil leer
Klasse Spam
Bemerkung Der Autor dieses Tweet postet Werbungen.

Tweet Claritin D on deck
Benutzername @DanielleMamath
Profil Rocky mountain goats
Klasse Other
Bemerkung Aus dem Inhalt ist zu unklar zu sehen, was dies Tweet bedeutet, es geht nicht um die Wirkung des Arzneimittels. Von dem Profil wird auch keine nützliche Information gezeigt. Deswegen gehört dies Tweet zu der Klasse *Other*

Tweet Autocorrect changed productivityto Prilosecïn that previous tweet. Fitting.
Benutzername @FarrahRochon
Profil USA Today Bestselling romance novelist, chocolate lover, sports junkie, Disney enthusiast and Broadway show fanatic.
Klasse Other
Bemerkung In diesem Tweet geht es inhaltlich nicht um das Arzneimittel.

3.5 Implementierung

Bevor die Suche in Twitter durchgeführt werden kann, werden die Synonyma der Medikamente aus DrugBank 2.2.3 gebraucht. Unter Verwendung von StAX (Streaming API for XML) können die Synonyma jedes Arzneimittels aus der XML-Datei[2] von DrugBank gelesen werden. Die ausgelesenen Ergebnisse werden in einer Hash-Tabelle im Json-Format gespeichert.

In der Phase der Anwendung von Twitter API werden alle Suchparameter in einer Java-Klasse verkapselt.

- **create:** Eine leere Instanz wird erstellt, in die die Suchparameter hinzugefügt werden.
- **setQuerySearch(searchName):** *searchName* ist eine Abfrage (Datentyp:String), mit der in Twitter gesucht werden kann. Die oben genannte, in Json-Format gespeicherte, Liste der Arztmittel, einschließlich der Synonyma, wird als Abfrage in Twitter zur Suche verwendet.
- **setSince("2008-12-01"):** Eine Datumsuntergrenze zur Beschränkung des Suchzeitraums;
- **setUntil("2015-11-30"):** Eine Datumsobergrenze zur Beschränkung des Suchzeitraums.

[2]Die Datei umfasst die Informationen der zugelassenen Medikamente 2015.
http://www.drugbank.ca/downloads# external-links

4 Trainieren des Klassifikators und Implementierung

Text-Kategorisierung ist der Prozess, der die Kategorie der Texte nach ihren Inhalten in die entsprechende vorgegebene Klasse automatisch einordnen kann. Er ist grundsätzlich in zwei Phasen unterteilt: Trainieren und Testen. In diesem Kapitel wird der Aufbau des Klassifikators ermittelt.

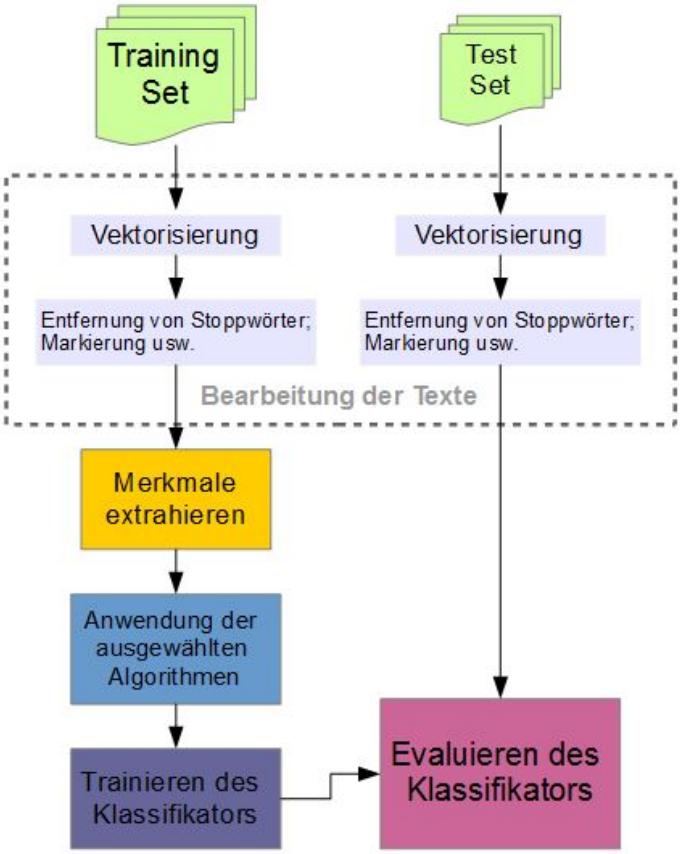


Abbildung 4.1: Verfahren zum Trainieren eines Klassifikator

4.1 Extrahieren der Merkmale

Das Corpus, das aus den normalisierten Tweets besteht, wird für das Extrahieren der Merkmale verwendet. Danach werden die Klassifikatoren trainiert. Chu et al. [Gam04] haben soziale Spams in Twitter detektiert. Bei der Auswahl der Merkmale legten sie ihre Aufmerksamkeit auf die in den Tweets enthaltenen URLs. Dazu berücksichtigten sie die Informationen der

Benutzerprofile und eine Liste der Stichwörter der häufigst benutzten Spamwörter. In der Studie von M. McCord und M. Chuah, die sich mit der Erkennung von Spams in Twitter befassten [MC11], extrahierten sie auch deren Merkmale unter Berücksichtigung der Benutzerinformationen und Textinhalt. Dazu verwerteten sie noch die Anzahl der Freunde („ FOLGE ICH"), Anzahl der Follower und den Ruf eines Benutzers, genau so wie die Länge der Tweets und ob URLs enthalten sind.

In dieser Arbeit wird mehr als eine Klasse verwendet, deshalb reicht allein die Analyse und Bearbeitung von URLs für das Extrahieren der Merkmale nicht aus. Benutzerorientierte Merkmale helfen hier auch nicht unbedingt, da die User aus den Klassen Patient, Arzt oder Forscher auch gewöhnliche Benutzer sind. Wir können nicht anhand der Anzahl an „FOLGE ICH" oder „FOLLOWER" erkennen, zu welcher Klasse ein Tweet gehört. Aus diesen Gründen wurden die Algorithmen Information Gain & Mutual Information, LDA Topic-Model und N-Gram eingeführt. Die Standardeinstellung für die Auswahl der Merkmale in Mallet ist, dass alle Wörter/Token/Terme in allen Dokumenten als Merkmale in der Form von unigram zur Kategorisierung genommen werden.

N-Gram

N-Gram Model [CT+94] basiert auf der Annahme, dass wenn das n-te Wort nur von seinem (n-1)-te Wort abhängig ist, heißt es Bigram (hier N=2). wenn das n-te Wort vom (n-2)-te und (n-3)-te Wörter abhängig ist, nennt man es Trigram (hier N=3). Mit einer mathematischer Formel ist es einfacher zu erklären:

$$P(S) = \prod_{i=1}^{n} P(w_i | \, w_1, w_2, ..., w_i - 1) \tag{4.1}$$

N-gram kann bei der Erfassung der gemein gebräuchlichen Phrasen und speziellen Wortgruppen sehr hilfreich sein. Allerdings je größere N ist, desto mehr muss berechnet werden und desto länger dauert die Berechnung. Deswegen splitten wir die Sätze nicht nur mit Unigram sondern auch mit 2-Gram und 3-Gram. Folgend wird eine Beispiel gezeigt:

Original: A great article on basil. Here is 'something to add' first: If you are taking coumadin as a blood-thinner,... URLXXX

Unigram: A great article on basil Here is something to add first If you are taking coumadin as a blood thinner URLXXX

Bigram: A_great great_article article_on on_basil basil_Here Here_is is_something something_to to_add add_first first_If If_you you_are are_taking taking_coumadin coumadin_as as_a a_blood blood_thinner thinner_URLXXX

Trigram: A_great_article great_article_on article_on_basil on_basil_Here basil_Here_is Here_is_something is_something_to something_to_add to_add_first add_first_If first_If_you If_you_are you_are_taking are_taking_coumadin taking_coumadin_as coumadin_as_a as_a_blood a_blood_thinner blood_thinner_URLXXX

Unter Bigram und Trigram werden die einzelnen Wörter durch das Zeichen „ _“ miteinander verbunden, wodurch ein „ neues Wort“ erzeugt wird. Beispielsweise ist es sinnvoll, aus den zwei Wörter „ blood “ und „ thinner “ nach Bigram ein neues „ blood_thinner“ zu erzeugen.

4.1.1 Entfernung der Stoppwörter

Mallet bietet eine Standard-Stoppwörter-Liste an. Die Liste enthält insgesamt 524 Wörter inklusive der Buchstaben a-z. Mit Kommandozeile *remove-stopwords* kann man bei Import der Dateien die Stopowörter gleichzeitig wegnehmen.

Abbildung 4.2: Beispiel der Stoppwörter auf Englisch[1]

4.1.2 Information Gain & Mutual Information in Mallet

Die Methode für die Auswahl der Merkmale anhand Infomation Gain & Mutual Information kann in Mallet auf eine einfache Weise, nämlich mit der Kommandozeile, erreicht werden. Durch die Kommandozeile *print-word-infogain* erhält man eine Liste der Wörter, die die höchsten durchschnittlichen Mutual Information ($I(U;C)$) mit den Klassenvariablen haben. Von dem ganzen Corpus dieser Arbeit werden insgesamt 1000 Merkmale anhand von InfoGain& MI ausgewählt. Die 1000 Merkmale werden fünf mal unterteilt. Bei ersten mal werden die Top 100, dann die Top 300, 500, 800 und dann alle Merkmale ausgewählt. In Abbildung 4.3 wird die Kommandozeile in Mallet hervorgehoben. Hier sieht man, dass mit dem Buchstaben N die Anzahl der Wörter gewählt wird.

[1] http://xpo6.com/list-of-english-stop-words/

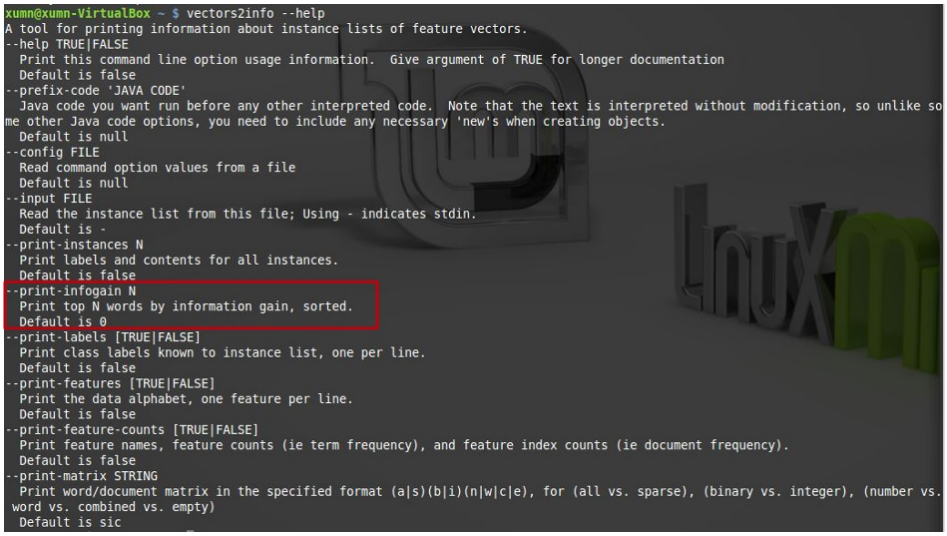


Abbildung 4.3: Kommandozeile in Mallet um die Wörterliste zu bekommen

4.1.3 LDA Topic Model

In dieser Arbeit wird das LDA Topic Model (folgend „LDA“gezeichnet) verwendet, mit dem die semantischen Themen jeder Klasse generiert werden. Durch die Reduktion der Dimensionen der Feature in den Dokumenten kann das LDA-Topic Model die Spärlichkeit der Dokumente deutlich verbessern [CHWE12]. Für jede Klasse werden jeweils 4 Mal die wahrscheinlichsten Themen generiert und jedes mal enthält das Thema eine jeweilige Anzahl von 100 bis 500 Wörter. Diese Wörter werden danach als Merkmale zur Klassifikation verwendet.

Mit dem Befehl *train-topics -- input tutorial.mallet* [2] kann die Datei (z.B tutorial.mallet) geöffnet werden. In Mallet lassen sich die Topics iterativ und automatisch trainieren [SGM]. Zusätzliche optionale Kommandozeilen:

- -num-top-words INTEGER Gibt die Anzahl der wahrscheinlichsten Wörter für jedes Thema nach Schätzung des Modells an.

- -num-topics INTEGER Gibt die Anzahl der Themen an, d.h wie viele Themen trainiert werden.

- -optimize-interval N Gibt die Anzahl der Iterationen zwischen Wiederschätzung von Dirichlet-Hyperparameter an, sodass das Modell besser zu den Dateien passt, indem es einige Themen bevorzugt behandelt. Mit 10 Mal Iterationen wird es logisch.

- -output-topic-keys Speichert die Top-Wörter für jedes Thema und entsprechende Dirichlet-Parameter.

[2] Die Datei der Form „.mallet“ existiert speziell in MALLET. Importieren die txt.Datei und MALLET exportiert mallet.Datei: mit Kommandozeile $import - dir - -input - -output$.

4.2 Klassifikator

In Kapitel 2 haben wir die Grundkenntnisse der zwei Klassifikatoren Maximum Entropy 2.3.1 und naiver Bayes 2.3.1 kurz eingeführt. In dieser Arbeit wird hauptsächlich der Maximum Entropy Klassifikator (im folgenden abgekürzt mit MaxEnt) verwendet. Die Ergebnisse anderer Klassifikatoren wie von naivem Bayes-Klassifikator und C4.5 Klassifikator (A C4.5 Decision Tree Classifier)[3] werden mit dem MaxEnt-Klassifikator verglichen.

Das Maximum Entropy Modell 2.3.1 ist ein statistisches Modell, das sich gut zur Lösung des Klassifikationsproblemes eignet. Mit Mallet können direkt durch die Kommandozeile der MaxEnt-Klassifikator oder gleichzeitig mehrere Klassifikatoren aufgerufen werden.

bin/mallet train-classifier - -input training.mallet - -output-classifier my.classifier
bin/mallet train-classifier - -input training.mallet - -output-classifier my.classifier - -trainer MaxEnt - -trainer NaiveBayes

davon:

train-classifier Startet das Trainieren des Klassifikators.

- -input Gibt die Training Set an;

- -output Gibt den trainierten Klassifikator aus.

- -trainer Wählt den Algorithmus zur Klassifikation aus. Es können auch mehrere Algorithmen gleichzeitig ausgewählt werden.

Mallet ist ein sehr gutes Tool für die Klassifikation. Es kapselt die komplexen Algorithmen und mathematische Formeln so, dass sie direkt verwendet werden können.

[3] Der C4.5 Entscheidungsbaum kann zur Klassifizierung und auch zur statistischer Klassifikation verwendet werden. Er basiert auf dem Konzept der „ Information Entropy 2.3.1“ durch maschinelles Lernen einen Entscheidungsbaum zu erstellen. Da C4.5 nicht das Studienobjekt dieser Arbeit ist, wird es nicht weiter erklärt.

5 Corpus Analyse

Die Verteilung der Top-Arzneimittel in der jeweiligen Klasse und die anhand der Methoden InfoGain & MI und LDA-Topic-Model ermittelten Top-Merkmale werden in diesem Kapitel analysiert.

5.1 Die vorkommenden Top-Arzneimittel

Das Ziel der Analyse des Corpus besteht darin, die Verteilung der Arzneimittel im Corpus darzustellen. Es wird also bestimmt, welche Arzneimittel am meisten von welchen Gruppen gepostet werden. Bei 130 Arzneimitteln und 520 entsprechenden Synonyma kommen 273 verschiedene Arzneimittelnamen im Corpus (insgesamt 2920 Tweets) vor. Die drei Arzneimittelnamen „Hydrocodone“, „Lipitor“ und „Omeprazole“ tauchen in allen acht Klassen auf. Die Arzneimittelnamen „Herceptin“, „Naloxone“ und „Olanzapine“ kamen in 7 Klassen (außer *Other*) vor. In der Klasse *Arzt* wurden 143 unterschiedliche Arzneimittelnamen erwähnt, in den anderen 7 Klassen jeweils 130 in *Forscher*, 131 in *Journalist*, 123 in *Spam*, 107 in *Patient*, 87 in *Behörden*, 77 in *Pharma* und 44 in *Other*.

Klassen	Arzneimittel	Anzahl	Rezept	Hauptanwendung [1]
Arzt	Azithromycin	23	Ja	Antibiotikum, Antiphlogistikum
	Warfarin	20	Ja	Antikoagulation, gegen Thrombose
	Herceptin	19	Nein	Brustkrebs und Magenkrebs
	Clopidogrel	15	Ja	Thrombozytenaggregationshemmer
	Furosemide	15	Ja	Bluthochdruck, Aszites und Ödeme
	Lidocaine	15	Ja	Aphthe, Lokalanästhesie, Antiarrhythmikum
	Ibuprofen	13	Ja	Schmerzen, Entzündungen und Fieber
	Naloxone	12	Ja	als Antidot bei Opiatüberdosierung
	Ativan	10	Ja	Angststörung
	Rituximab	10	Ja	Non-Hodgkin-Lymphom, Rheumatoide Arthritis
	Testosterone	10	Ja	Hypogonadismus

Tabelle 5.1: Die Top-Arzneinmittelnamen in der Klasse *Arzt*

Die Tabellen 5.1 bis 5.8 zeigen wie häufig die Arzneimittel in der jeweiligen Klasse vorkommen. Für die Klassen *Behörden*, *Pharma* und *Other* mit relative wenig Tweets in den Trainingsdateien, wurden nur die Arzneimittel gelistet, die mindestens 5 mal vorgekommen sind. Für die anderen 5 Klassen wurden nur die Arzneimittel gelistet, die mindestens 10 mal vorgekommen sind. In der Tabelle mit den Top-Arzneimittelnamen enthalten die Klassen *Arzt* und *Patient* drei gemeinsame Arzneimittelnamen: *Ibuprofen, Lidocaine und Ativan*. Diese drei Arzneimittel behandeln Volkskrankheiten wie leichte Entzündung, Schmerzen, Fieber, Angststörung und Antiarrhythmikum. Die Klasse *Arzt* hat ansonsten nur jeweils einen gleichen Arzneimittelnamen wie die Klassen *Journalist* und *Behörden*. Die Top-Arzneimittel der

Klasse *Forscher* betrifft vor allem die Behandlung von Tumoren oder Krebs. In Klasse *Behörden* tauchen einige Synonyma auf. *Pegfilgrastim, Neulasta und g-CSF* sind dasselbe Medikament und *Insulin und Lantus* sind auch dasselbe Arzneimittel zur Absenkung des Blutzuckerspiegels. In der Klasse *Patient* kamen ausnahmsweise keine Antikrebs-/Antitumor-Medikamente vor. Die Arzneimittel, die in der Klasse *Patient* diskutiert wurden, sind zur Behandelung von Schmerzen, Magen-und Darmkrankheiten und psychischer Störungen. Die in der Klasse *Spam* vorkommende Arzneimittel tauchen fast alle in anderen Klassen auch auf. Die Klassen *Spam* und *Other* haben fünf gleiche Arzneimittelnamen. Von den Top-Arzneimitelle sind die verschreibungspflichtigen Arzneimittel in der Überzahl.

Klassen	Arzneimittel	Anzahl	Rezept	Hauptanwendung
Forscher	Bevacizumab	50	Ja	Behandlung von sechs fortgeschrittenen Krebserkrankungen (Darm-, Lungen-, Brust-, Nieren-, Eierstock- und Gebärmutterhalskrebs)
	Trastuzumab	37	Nein	Brustkrebs und Magenkrebs
	Cepecitabine	21	Ja	Therapie von metastasiertem Dickdarmkrebs, metastasiertem oder lokal fortgeschrittenem Mammakarzinom
	Avastin	16	Ja	Behandlung von sechs fortgeschrittenen Krebserkrankungen (Darm-, Lungen-, Brust-, Nieren-, Eierstock- und Gebärmutterhalskrebs)
	Imatinib	14	Ja	Behandlung der chronischen myeloischen Leukämie (CML), von gastrointestinalen Stromatumoren (GIST) sowie weiteren malignen Erkrankungen
	Niacin	12	Nein	Absenkung erhöhter Blutfettwerte und LDL-Cholesterinn, Erhöhung des HDL-Cholesterin

Tabelle 5.2: Die Top-Arzneinmittelnamen in der Klasse *Forscher*

Klassen	Arzneimittel	Anzahl	Rezept	Hauptanwendung
Pharma	Claritin	44	Nein	Linderung der Beschwerden bei Allergien und beim atopischen Ekzem (Neurodermitis)
	Amoxicillin	11	Ja	Behandlung von Infektionen
	Pegfilgrastim	11	Ja	Chemotherapie-induzierter Leukopenie
	Adalimuab	10	Ja	Behandlung von rheumatoider Arthritis, Psoriasis-Arthritis, Spondylitis ankylosans und der chronisch entzündlichen Darmerkrankungen Morbus Crohn und Colitis ulcerosa
	Etanercept	8	Ja	Behandlung rheumatischer Erkrankungen und der Psoriasis
	Neulasta	7	Ja	Chemotherapie-induzierter Leukopenie
	Xeloda	7	Ja	Therapie von metastasiertem Dickdarmkrebs, metastasiertem oder lokal fortgeschrittenem Mammakarzinom
	g-CSF	7	Ja	Chemotherapie-induzierter Leukopenie
	Avastin	6	Ja	Behandlung von sechs fortgeschrittenen Krebserkrankungen (Darm-, Lungen-, Brust-, Nieren-, Eierstock- und Gebärmutterhalskrebs)
	Insulin	6	Nein	Absenkung des Blutzuckerspiegel
	Lantus	6	Ja	Absenkung des Blutzuckerspiegel
	Albuterol	5	Ja	als Bronchospasmolytikum bei Asthma bronchiale
	Combivent	5	Ja	Behandlung von chronisch obstruktiven Lungenerkrankungen und Herzrhythmusstörungen
	Ipratropium bromide	5	Ja	Behandlung von chronisch obstruktiven Lungenerkrankungen und Herzrhythmusstörungen
	Prevnar 13	5	Ja	gegen die Haupterreger der infektiösen, bakteriellen Pneumonie, die Pneumokokken
	Tiotropium	5	Ja	Chronisch obstruktive Lungenerkrankung (COPD)

Tabelle 5.3: Die Top-Arzneinmittelnamen in der Klasse *Pharma*

Klassen	Arzneimittel	Anzahl	Rezept	Hauptanwendung
Journalist	Abilify	23	Ja	Behandlung der Schizophrenie
	Enbrel	21	Nein	Behandlung rheumatischer Erkrankungen und der Psoriasis
	Hydrocodone	21	Ja	Analgetikum (Schmerzmittel), Antitussivum (Hustenmittel)
	Insulin	16	Nein	Absenkung des Blutzuckerspiegel
	Velcade	15	Ja	Therapie des multiplen Myeloms (Plasmozytom)
	Lucentis	14	Nein	Behandlung der feuchten (exsudativen) altersbezogenen Makuladegeneration (AMD)
	Niacin	14	Nein	Absenkung erhöhter Blutfettwerte und LDL-Cholesterinn, Erhöhung des HDL-Cholesterin
	Imatinib	13	Ja	Behandlung der chronischen myeloischen Leukämie (CML), von gastrointestinalen Stromatumoren (GIST) sowie weiteren malignen Erkrankungen
	Ativan	12	Ja	Angststörung
	Cymbalta	12	Ja	Behandlung von Depressionen, generalisierten Angststörungen, diabetischer Polyneuropathie und Harninkontinenz
	Januvia	12	Ja	Behandlung des Diabetes mellitus Typ 2
	Azithromycin	10	Ja	Antibiotikum, Antiphlogistikum
	Bevacizumab	10	Ja	Behandlung von sechs fortgeschrittenen Krebserkrankungen (Darm-, Lungen-, Brust-, Nieren-, Eierstock- und Gebärmutterhalskrebs)
	Metformin	10	Ja	Antidiabetikum

Tabelle 5.4: Die Top-Arzneinmittelnamen in der Klasse *Journalist*

Klassen	Arzneimittel	Anzahl	Rezept	Hauptanwendung
Patient	Hydrocodone	31	Ja	Analgetikum (Schmerzmittel), Antitussivum (Hustenmittel)
	Vyvanse	26	Ja	Behandlung von ADHS und Narkolepsie
	Ibuprofen	23	Ja	Schmerzen, Entzündungen und Fieber
	Ativan	17	Ja	Angststörung
	Prilosec	15	Ja	Behandlung von Magen- und Zwölffingerdarmgeschwüren sowie bei Refluxösophagitis
	Vitamin D	15	Nein	Nahrungsergänzungsmittel
	Lidocaine	14	Ja	Aphthe, Lokalanästhesie, Antiarrhythmikum
	Ibuprophen	13	Ja	Schmerzen, Entzündungen und Fieber
	Adderall	12	Ja	Behandlung von ADHS und Narkolepsie
	Cymbalta	12	Ja	Behandlung von Depressionen, generalisierten Angststörungen, diabetischer Polyneuropathie und Harninkontinenz
	Tramadol	11	Ja	Behandlung mäßig starker bis starker Schmerzen

Tabelle 5.5: Die Top-Arzneinmittelnamen in der Klasse *Patient*

Klassen	Arzneimittel	Anzahl	Rezept	Hauptanwendung
Behörden	Naloxone	29	Ja	als Antidot bei Opiatüberdosierung
	Metformin	10	Ja	Antidiabetikum
	T4	9	Nein	Schilddrüsenhormone
	Pemetrexed	8	Ja	Zytostatikum (Antimetabolit)
	Hydrocodone	6	Ja	Analgetikum (Schmerzmittel), Antitussivum (Hustenmittel)
	Avastin	6	Ja	Behandlung von sechs fortgeschrittenen Krebserkrankungen (Darm-, Lungen-, Brust-, Nieren-, Eierstock- und Gebärmutterhalskrebs)
	Clonazepam	5	Ja	Antiepileptikum, REM-Schlaf-Verhaltensstörung, Epilepsie

Tabelle 5.6: Die Top Arzneinmittelnamen in der Klasse *Behörden*

Klassen	Arzneimittel	Anzahl	Rezept	Hauptanwendung
Spam	Prilosec	44	Ja	Behandlung von Magen- und Zwölffingerdarmgeschwüren sowie bei Refluxösophagitis
	Abilify	30	Ja	Behandlung der Schizophrenie
	Azithromycin	29	Ja	Antibiotikum, Antiphlogistikum
	Gabapentin	19	Ja	Behandlung der Epilepsie und neuropathischer Schmerzen
	Neurontin	19	Ja	Behandlung der Epilepsie und neuropathischer Schmerzen
	Nexium	18	Ja	Behandlung von Magen- und Zwölffingerdarmgeschwüren sowie bei Refluxösophagitis (einer entzündlichen Erkrankung der Speiseröhre)
	Celebrex	15	Ja	Behandlung von degenerativen Gelenkerkrankungen, chronischer Polyarthritis und Morbus Bechterew
	Claritin	14	Nein	Linderung der Beschwerden bei Allergien und beim atopischen Ekzem (Neurodermitis)
	Cephalexin	12	Ja	Behandlung von bakteriellen Infektionen im Bereich der Harn- und Geschlechtsorgane, der Atemwege, der Haut und des Weichteilgewebes, des Hals-, Nasen- und Ohrengebietes, der Knochen und Gelenke sowie der Zähne
	Zithromax	12	Ja	Antibiotiku,Antiphlogistikum
	Lisinopril	11	Ja	Behandlung der arteriellen Hypertonie (Bluthochdruck) und der Herzinsuffizienz
	Venlafaxine	11	Ja	Behandlung von Depressionen und Angsterkrankungen
	Lipitor	10	Ja	Therapie der Hypercholesterinämie
	Niacin	10	Nein	Absenkung erhöhter Blutfettwerte und LDL-Cholesterinn, Erhöhung des HDL-Cholesterin

Tabelle 5.7: Die Top-Arzneinmittelnamen in der Klasse *Spam*

Klassen	Arzneimittel	Anzahl	Rezept	Hauptanwendung
Other	Abilify	19	Ja	Behandlung der Schizophrenie
	Claritin	11	Nein	Linderung der Beschwerden bei Allergien und beim atopischen Ekzem
	Prilosec	9	Ja	Behandlung von Magen- und Zwölffingerdarmgeschwüren sowie bei Refluxösophagitis
	Niacin	8	Nein	Absenkung erhöhter Blutfettwerte und LDL-Cholesterinn, Erhöhung des HDL-Cholesterin
	Vyvanse	8	Ja	Behandlung von ADHS und Narkolepsie
	Hydrocodone	6	Ja	Analgetikum (Schmerzmittel), Antitussivum (Hustenmittel)
	Xanax	6	Ja	Behandlung von Angst- und Panikstörungen
	Azithromycin	5	Ja	Antibiotikum, Antiphlogistikum
	Opana	5	Ja	Akut- und Langzeit-Schmerztherapie bei starken bis sehr starken Schmerzen

Tabelle 5.8: Die Top-Arzneinmittelnamen in der Klasse *Other*

In der Tabelle 5.9 werden die Top-20 Arzneimittelnamen gezeigt, die am meisten im Corpus vorgekommen sind. Abgesehen davon, dass die Verteilung der Arzneimittelnamen in den Klassen *Pharma* und *Other* eine Spärlichkeit hat, kommen die Top-20 Arzneimittelnamen in fast jeder Klasse vor. Das zeigt, dass diese Top-20 Arzneimittelnamen unter den acht sozialen Gruppen relativ oft erwähnt und diskutiert werden. Die Top-Arzneimittel werden hauptsächlich für die Behandlungen von Schmerzen, Krebs/Tumor, psychischer Störung, Antibiotikum, Diabetes und Herz-Kreislauf-Krankheiten sowie Allergien, eingesetzt.

	Arzt	Behörden	Forscher	Journalist	Other	Patient	Pharma	Spam	Sum
Abilify			7	23	19	7		30	86
Hydrocodone	4	6	6	21	6	31	1	6	81
Azithromycin	23	3	4	10	5	5		29	79
Prilosec	6	2		3	9	15		44	79
Claritin	2		2		11		44	14	73
Bevacizumab	2	2	50	10			1		66
Naloxone	12	29	4	3		6	2	3	59
Niacin	3		12	14	8	7		10	54
Trastuzumab	7	3	37	5			2		54
Avastin	8	6	16	12			6		48
Vyvanse	3		4	6	8	26		1	48
Gabapentin	3	2	3	9		8		19	44
Ibuprofen	13	1	3	1		23		3	44
Lidocaine	15	4	1	3		14		7	44
Lipitor	3	1	8	7	4	7	1	10	41
Nexium	4		3	7	2	3	3	18	40
Metformin	7	10		10		5	1	5	38
Clopidogrel	15	5	6	9			2		37
Ativan	10					17		7	34
Capecitabine	8	2	21	2					33
Cymbalta	1		2	12	1	12		5	33
Herceptin	19	3	5	3	1	1	1		33
Venlafaxine	1	2	1	7	1	9		11	32
Warfarin	20	3	6	2			1		32
Imatinib	2	2	14	13					31

Tabelle 5.9: Die Top-20 Arzneinmittelnamen im Corpus

5.2 Top-Wörter anhand Infomation Gain

Unter der Verwendung von InfoGain& MI wurden die acht Klassen als Ganzes betrachtet und mit ihnen wurden die Top Merkmale bestimmt (nicht je nach Klasse die eigene Merkmale). Die Tabelle 5.10 zeigt die Top-30-Wörter der acht Klassen.

Top 30 Wörter aus InfoGain, sortiert nach Mutual Information									
1	fda	11	patients	21	patent	31	order	41	learn
2	username	12	prilosec	22	i'm	32	news	42	presented
3	buy	13	generic	23	cheap	33	overdose	43	launch
4	online	14	sandoz	24	trastuzumab	34	approves	44	recall
5	drugname	15	naloxone	25	approval	35	risk	45	free
6	phase	16	data	26	communication	36	drug	46	treatment
7	bevacizumab	17	study	27	commercial	37	pts	47	mylan
8	claritin	18	abilify	28	trial	38	capecitabine	48	enbrel
9	biosimilar	19	cancer	29	launches	39	breast	49	investigational
10	safety	20	pharma	30	therapy	40	announces	50	u.s

Tabelle 5.10: Top 30 Wörter des ganzen Corpus anhang von InfoGain, sortiert nach Mutual Information

5.3 Die wahrscheinlichsten Themen jeder Klasse anhand LDA-Topic-Model

Unter der Verwendung des LDA-Topic-Models wurden die wahrscheinlichsten Themen jeweiliger Klassen bestimmt. Die Tabelle 5.11 zeigt die acht wahrscheinlichsten Themen jeder Klasse, die jeweils 30 Top-Wörter enthalten.

Klasse	Das wahrscheinlichste Thema mit 30 Wörter
Arzt	username urlxxx patients drug study pts azithromycin warfarin risk cancer therapy herceptin foamed urlpic furosemide patient it's clopidogrel lidocaine treatment high evidence dose give pain testosterone low good naloxone ibuprofen
Behörden	urlxxx fda username drug naloxone safety cancer overdose risk communication recall approves treatment urlpic drugs patients products lives pemetrexed avastin access deaths label factor breast abcdrbchat hydrocodone jco nsclc review
Forscher	urlxxx username bevacizumab cancer study patients trastuzumab phase trial treatment breast drug capecitabine therapy oncology risk pharma avastin adjuvant chemotherapy hiv fda imatinib positive combination niacin advanced drugs iii metformin
Journalist	urlxxx fda drug patent news pharma bioportfolionews patients study treatment drugs cancer enbrel research generic risk abilify approves court username hydrocodone biosimilar combination trial markets u.s insulin lucentis approval merck
Patient	username claritin i'm hydrocodone vyvanse ibuprofen sleep vitamin i've today feel niacin time pain good taking night back don't ativan day prilosec work it's sunshine can't lidocaine lol ibuprophen doctor
Pharma	urlxxx username data biosimilar fda sandoz phase approval generic launches tablets treatment presented learn launch announces trial mylan patients asthma today iii investigational biosimilars amoxicillin asco pegfilgrastim actavis adalimumab capsules
Spam	urlxxx buy online username drugname generic prilosec cheap order abilify urlpic azithromycin side free prescription effects tablets drug gabapentin neurontin viagra nexium pharmacy celebrex claritin otc cephalexin health tramadol zithromax
Other	username abilify commercial claritin viagra urlxxx prilosec guy love niacin xanax vyvanse day drugname hydrocodone azithromycin larry don't lipitor januvia it's movie woman people cable that's oscars shot sprite friday

Tabelle 5.11: Top 30 Wörter des ganzen Corpus anhang von InfoGain, sortiert nach Mutual Information

6 Evaluation

In diesem Kapitel werden zuerst die Ergebnisse der verschiedenen Klassifikatoren und verschiedenen Merkmalen erklärt. Außerdem legen wir die zwei Klassen, *Arzt* und *Forscher*, zusammen und vergleichen das neue Ergebnis mit den Ergebnissen aller Klassen.

6.1 Vergleiche zwischen Klassifikatoren und Merkmalen

Die Klassifikatoren in Mallet verwenden alle Feature (Wörter) des Dokuments als Merkmale zur Kategorisierung. Für kurze Texte und kleine Training Sets ist es immer noch durchführbar aber langsam. Im Fall von mehreren und großen Dokumenten ist die Dimension der Merkmalsvektoren sehr groß. Wenn alle Wörter als Merkmale verwendet werden, werden die Kosten und Zeiten der Berechnung sehr groß und fast unberechenbar. Eine weitere Klassifikation ist in diesem Fall fast unmöglich. Da die Datei in dieser Arbeit nicht zu groß ist, wird das Ergebnis von dem mit allen Feature als Bezug verwendet.
In Kapitel 2.3.2 werden drei Methoden zur Evaluation eines Klassifikators beschrieben. Daher werden die Ergebnisse von verschiedenen Klassifikatoren mit unterschiedlichen Merkmalen durch Berechnung ihrer F1-Measure nach 10-Cross-Validation bestimmt.

Klassifikatoren: Maximum Entropy VS. naiver Bayes

Die F1-Measure in Abbildung 6.1 zeigen deutlich, dass naiver Bayes-Klassifikator im Vergleich zu MaxEnt Klassifikator, bei kleinen Training Sets, wie den Klassen *Behörden* und *Other*, besser abschneiden. Bei anderen zeigt der naive Bayes-Klassifikator auch keine herausstechend gute Fähigkeit. Um gut ausgewogene Ergebnisse zu erhalten, wird hauptsächlich der Maximum Entropy Klassifikator verwendet.

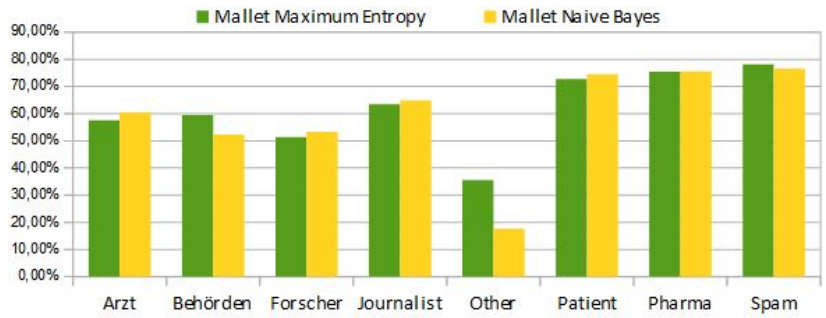


Abbildung 6.1: Vergleich der Performance zwischen den Klassifikatoren Max Entropy und Naive Bayes

Merkmale: InfoGain (sortiert bei Mutual Information) und LDA-Topic-Model

In Kapitel 2.3.1 und 4.1 wurden die Methoden zum Extrahieren der Merkmale vorgestellt. Mit kurzen Worten kann man die zwei Hauptideen zusammenfassen. Die erste ist, die Top-Wörter aus jeweils dem Algorithmus InfoGain & MI und dem LDA-Topic-Model als Merkmale zu verwenden. Die zweite ist, die Merkmale aus InfoGain durch das LDA-Topic-Model zu ergänzen. Tabelle 6.1 zeigt die F1-Measure vom MaxEnt mit unterschiedlichen Merkmalen. In der Klasse *Patient* ähneln sich die F1-Measure, ihr Unterschied beträgt nur $\pm 0,84\%$. Die Ergebnisse der Methoden *InfoGain200*, *LDA200* und *InfoGain100+LDA100* heben sich etwas ab (in der Tabelle grün markierte Zellen). Mit *InfoGain200* und *LDA200* hat sich die F1-Measure in der Klasse *Arzt* um 3% und in der Klasse *Forscher* um ungefähr 1.5% verbessert. Mit *LDA200* und *InfoGain100+100LDA* haben sich die F1-Measure der Klasse *Journalist* um 3% verbessert. Abbildung 6.2 zeigt, dass die F1-Measure, außer in den Klassen *Behörden* und *Other*, sich in Bezug auf Mallet in allen anderen Klassen in unterschiedlichem Maße verbessert haben.

F1-Measure	Arzt	Behörden	Forscher	Journalist	Other	Patient	Pharma	Spam
Alle Feature[1]	57,5629%	59,4304%	51,3348%	63,4159%	35,5572%	72,7659%	75,4656%	78,1889%
InfoGain100[2]	58,9222%	58,8156%	51,7638%	63,5029%	35,4147%	72,8372%	76,7852%	78,1889%
InfoGain200	60,0432%	58,4787%	53,0138%	65,5405%	36,6161%	72,0871%	78,0194%	79,2565%
LDA100[3]	58,1982%	58,7293%	51,2479%	64,4835%	32,3706%	72,0621%	76,3827%	78,3949%
LDA200	60,7544%	60,7975%	52,7697%	66,4967%	34,6915%	72,127%	76,3236%	79,192%
InfoGain50 +LDA50[4]	59,8919%	58,0203%	51,0007%	65,6643%	34,9041%	72,0029%	77,1368%	78,9809%
InfoGain100 +LDA100	59,0617%	61,7567%	50,8982%	66,5721%	36,7831%	72,0656%	77,6393%	78,5786%

Tabelle 6.1: F1-Measure von MaxEnt mit verschiedenen Merkmalen

Für die Methoden InfoGain & MI und LDA-Topic-Model muss man immer bestimmen wie viele Merkmale ausgewählt werden sollen. Daher werden in dieser Arbeit verschiedene Anzahlen an Top-Wörter für jede Methode getestet. Im oberen Teil der Tabelle 6.2 wird ersichtlich, dass es relativ optimal ist, die Top-200 Wörter als Merkmalen unter der Verwendung von InfoGain& MI zu benutzen. Dasselbe gilt für die Verwendung von LDA-Topic-Model. Im unteren Teil der Tabelle werden die F1-Measure unter der Verwendung einer Kombination von InfoGain und LDA-Topic-Model verglichen. Dabei wird ersichtlich, dass die Verwendung von *InfoGain100 + LDA100* zu besseren Ergebnissen führt.

Bei den Methoden *InfoGain200, LDA200 und InfoGain100+LDA100* lässt sich nicht bestimmen, welche von ihnen die besten Ergebnisse liefert. Wie die gefärbten Zellen in der Tabelle zeigen führen alle drei Methoden für die Klassen *Journalist* und *Pharma* zu guten Resultaten. Die

[4]Alle Feature werden als Merkmale zum Klassifizieren verwendet.
[4]Nehmen Top 100 Wörter aus Training Set mit Methodik InfoGain und Mutual Information zum klassifizieren.
[4]Nehmen Top 100 Wörter der wahrscheinlichsten Themen aus Training Set mit Methodik LDA Topic Model zum klassifizieren.
[4]Nehmen jeweils Top 50 Wörter aus Training Set mit Methodik InfoGain & Mutual Information und LDA zum klassifizieren.

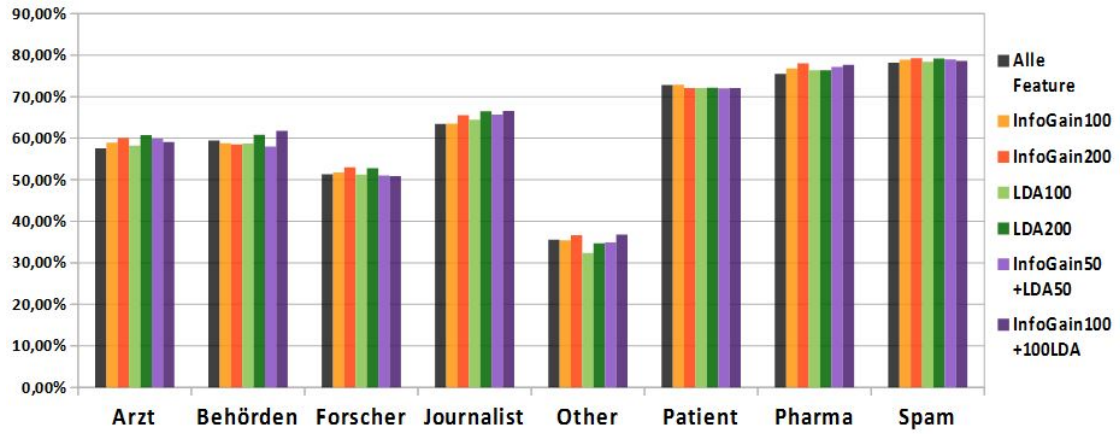


Abbildung 6.2: Vergleich der F1-Measure von MaxEnt mit verschiedenen Merkmalen

Kombination *InfoGain100+LDA100* funktioniert für die Klasse *Behörden* im Vergleich zu den anderen zwei Methoden am besten, bei den Klassen *Arzt* und *Spam* dagegen nicht. Der Grund dafür könnte sein, dass einige Wörter sowohl in *InfoGain100* als auch in *LDA100* vorkommen. Somit sinkt die Gesamtzahl der Top-Wörter im Gegensatz zu den Methoden *InfoGain200* und *LDA200*. LDA Topic Model verbessert die F1-Measure der Klasse *Other* nicht, weil die Inhalte der Tweets oft keine relevanten Aussagen über ein Medikament enthalten.

	Arzt	Behörden	Forscher	Journalist	Other	Patient	Pharma	Spam
Alle Feature	57,5629%	59,4304%	51,3348%	63,4159%	35,5572%	72,7659%	75,4656%	78,1889%
InfoGain100	58,9222%	58,8156%	51,7638%	63,5029%	35,4147%	72,8372%	76,7852%	78,9163%
InfoGain200	60,0432%	58,4787%	53,0138%	65,5405%	36,6161%	72,0871%	78,0194%	79,2565%
InfoGain300	57,5250%	60,1301%	51,4597%	64,5296%	36,2534%	72,4564%	77,0018%	78,2176%
InfoGain500	58,7652%	59,9260%	51,0214%	64,7530%	33,6858%	72,5032%	75,9551%	78,8022%
InfoGain800	58,7272%	58,1347%	52,9848%	64,0874%	34,2774%	73,0044%	76,9086%	77,4202%
InfoGain1000	58,0863%	59,7376%	51,4468%	63,9709%	29,6348%	71,9264%	76,6339%	78,3446%
Alle Feature	57,5629%	59,4304%	51,3348%	63,4159%	35,5572%	72,7659%	75,4656%	78,1889%
LDA100	58,1982%	58,7293%	51,2479%	63,4835%	32,4706%	72,0621%	76,3827%	78,3949%
LDA200	60,7544%	60,7976%	52,7698%	66,4967%	34,6916%	72,1270%	76,3236%	79,1920%
LDA300	59,1121%	60,0906%	51,4962%	65,0687%	31,3490%	71,9657%	75,3934%	79,33643%
LDA500	58,1734%	60,1466%	51,0810%	63,3444%	33,3455%	71,9621%	74,6861%	77,7111%
Alle Feature	57,5629%	59,4304%	51,3348%	63,4159%	35,5572%	72,7659%	75,4656%	78,1889%
InfoGain50 +LDA50	59,8919%	58,0203%	51,0007%	65,6643%	34,9041%	72,0029%	77,1368%	78,9809%
InfoGain100 +LDA30	58,1443%	58,1866%	52,563%	64,3304%	35,6715%	71,8552%	75,8039%	78,9589%
InfoGain100 +LDA50	59,0298%	60,8157%	49,2802%	64,1675%	31,1659%	72,294%	76,2227%	78,2807%
InfoGain100 +LDA100	59,0617%	61,7567%	50,8982%	66,5721%	36,7831%	72,0656%	77,6393%	78,5786%
InfoGain200 +LDA100	58,5933%	59,34402%	49,6378%	63,0501%	34,4873%	72,9128%	77,0785%	78,0599%

Tabelle 6.2: F1-Measure von InfoGain & MI und LDA-Topic-Model mit verschiedener Anzahl an Merkmale

Klassifikatoren: 2-Gram VS. 3-Gram

Um zu testen, ob 2-Gram und 3-Gram bei kurzen Texten auch gut funktionieren können, wird das Corpus jeweils in 2-Gram und 3-Gram umformuliert. Die Ergebnisse des Vergleichs der jeweils mit *InfoGain200* plus N-Gram, *LDA200* plus N-Gram und *InfoGain100+LDA100* plus N-Gramm entwickelten Klassifikatoren wird in Abbildung 6.3 gezeigt. Die Abbildung unten rechts zeigt das Resultat des Vergleichs mit der Standardmethode „Mallet“, wobei alle Token als Merkmale zur Klassifikation benutzt wurden. In diesem Fall verschlechtert die Methode N-Gram $(N = 2oder3)$ das Ergebnis enorm.

Auch die Ergebnisse der anderen drei Diagramme in der Abbildung zeigen, dass die Klassifikatoren mit der Methode N-Gram keine verbesserten F1-Measure liefern können. Die Methode N-Gram kann bei den kurzen Texten die Klassifikation nicht verbessern.

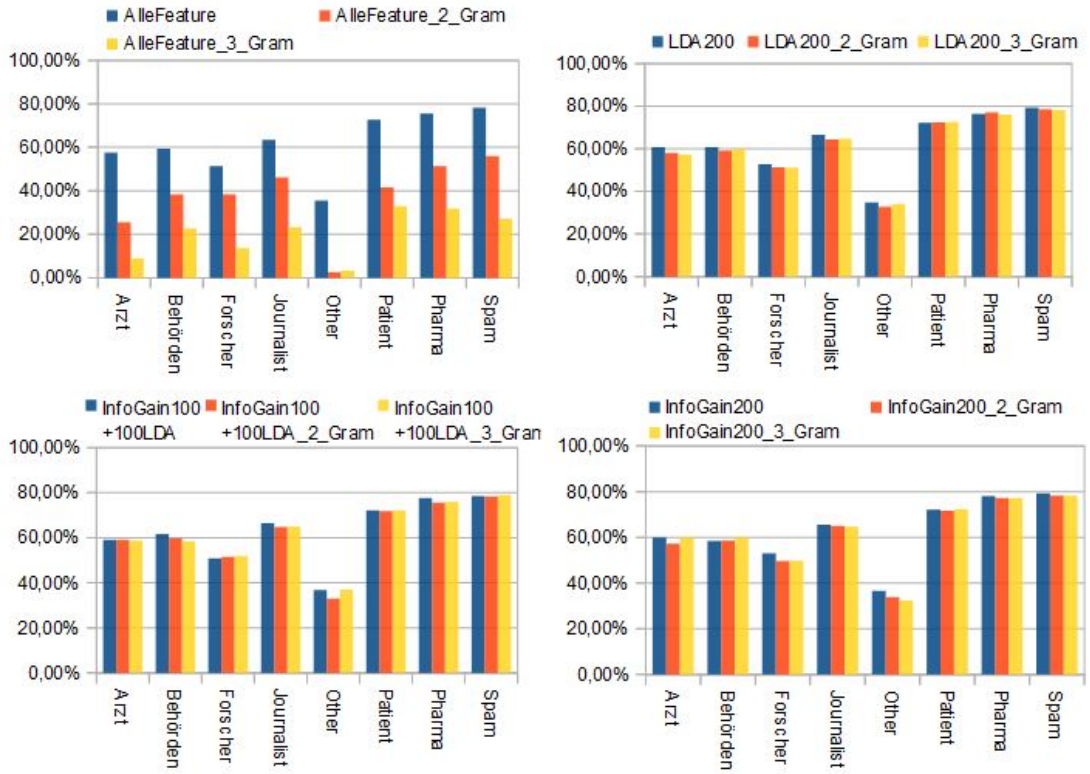


Abbildung 6.3: Vergleich der F1-Measure zwischen 2-Gram und 3-Gram mit verschiedenen Merkmalen

6.2 Verbinden der Klasse *Arzt* und *Forscher*

Die Tweets der Klassen *Arzt* und *Forscher* sind inhaltlich sehr ähnlich. Die Rolle der Benutzer dieser zwei Klassen ist nicht immer eindeutig. Ein Arzt z.B kann auch ein medizinischer Pro-

fessor sein. Ein Forscher kann auch ein fachlicher Mitarbeiter eines Instituts sein. Dazu wurde eine Konfusionsmatrix (siehe Tabelle 6.3) aus 10-Cross-Validation zufällig ausgewählt. In der Matrix werden 6 Tweets der Klasse *Arzt* falsch der Klasse *Forscher* zugeordnet und 7 Tweets aus *Forscher* der Klasse *Arzt*. Daher wurden die zwei Klassen in einer neuen Klasse **„AF“** zusammengefasst und die neue Klasse wurde mit den anderen 6 Klassen zur Kategorisierung verwendet.

label	Arzt	Behoerden	Forscher	Journalist	Other	Patient	Pharma	Spam	total
Arzt	26	1	6	2	.	6	.	5	46
Behoerden	1	8	1	4	.	.	1	2	17
Forscher	7	.	20	2	.	4	1	2	36
Journalist	4	.	4	39	.	1	.	7	55
Other	.	.	.	.	2	6	.	.	8
Patient	2	.	2	1	.	39	.	.	44
Pharma	.	.	1	3	.	1	16	.	21
Spam	1	.	.	.	.	.	1	34	36

Tabelle 6.3: Beispiel: Konfusionsmatrix

In dem in Tabelle 6.2 abgebildeten Vergleich liefern die Klassifikatoren mit den Methoden „InfoGain200“,„LDA200“ und „InfoGain100+LDA100“ bessere F1-Measure. Deshalb wurden die drei Methoden für die weitere Kategorisierung verwendet. Die Ergebnisse der F1-Measure werden in der Tabelle 6.4 gezeigt:

	AF	Behörden	Journalist	Other	Patient	Pharma	Spam
Alle Feature	73,6815%	55,6125%	62,6273%	30,8709%	73,9163%	76,6267%	79,2152%
InfoGain200	73,6593%	55,8746%	63,0813%	30,6737%	73,6936%	76,3385%	80,0902%
LDA200	73,7591%	54,3156%	63,6926%	29,1107%	73,5425%	76,2369%	80,0524%
InfoGain100 +LDA100	73,5897%	59,6842%	62,8531%	34,3437%	73,7515%	76,313%	79,3937%

Tabelle 6.4: F1-Measure nach der Verbindung der Klassen *Arzt* und *Forscher*

In einem Vergleich der Ergebnisse in Tabelle 6.2 und 6.4 wird ersichtlich, dass die F1-Measure der neuen Klasse **„AF“** im Vergleich zu den separaten *Arzt* und *Forscher* um 12% — 20% angestiegen sind. Im Gegensatz zum Ergebnis der Klasse **„AF“** verringert die F1-Measure der Klasse *Behörden*, unter der Verwendung von *InfoGain200* und *LDA200*, das Ergebnis um etwa 5%. Die Ergebnisse anderer Klassen schwanken um ungefähr $\pm 1\%$.

6.2.1 Klassifikation zwischen den Klassen *Arzt* und *Forscher*

Nach der Verbindung der Klassen *Arzt* und *Forscher* wird ein binärer Klassifikator gebraucht, um die Tweets diesen Klassen zuordnen zu können. In diesem Abschnitt wurden zuerst drei Klassifikatoren (Naive Bayes, MaxEnt und C45 4.2), mit der Standardeinstellung von Mallet, verglichen. In der Abbildung 6.4 wird ersichtlich, dass Naive Bayes in diesem Fall deutlich besser als die anderen zwei Klassifikatoren ist. Ausgehend von der Berücksichtigung

der Ähnlichkeit der Themen der Klassen *Arzt* und *Forscher* wurde als nächstes nur der Naiv Bayes Klassifikator mit der Methode „InfoGain& MI“, und nicht das LDA-Topic-Model, entwickelt.

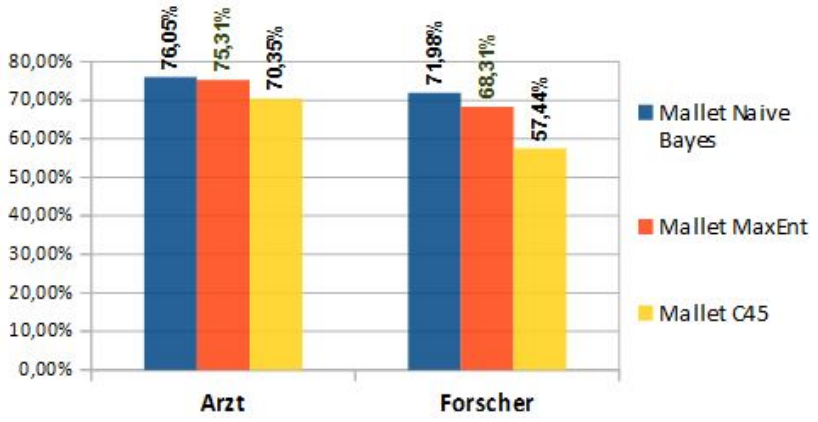


Abbildung 6.4: Vergleich verschiedener Klassifikatoren anhand „InfoGain& MI“

Um die Anzahl der Auswahl der Top-Wörter zu bestimmen wurden hierfür acht Klassifikatoren mit der Methode InfoGain entwickelt. Jeder Klassifikator hat eine verschiedene Anzahl an Merkmalen (nämlich Top-Wörter). Die Abbildung 6.5 zeigt als Ergebnis, dass eine erhöhte Anzahl an Top-Wörter den Klassifikator nur gering beeinflussen. Der Durchschnitt der F1-Measure der Klassen *Arzt* und *Forscher* sind jeweils 76,27% und 72,58%. Somit haben sich die F1-Measure deutlich verbessert.

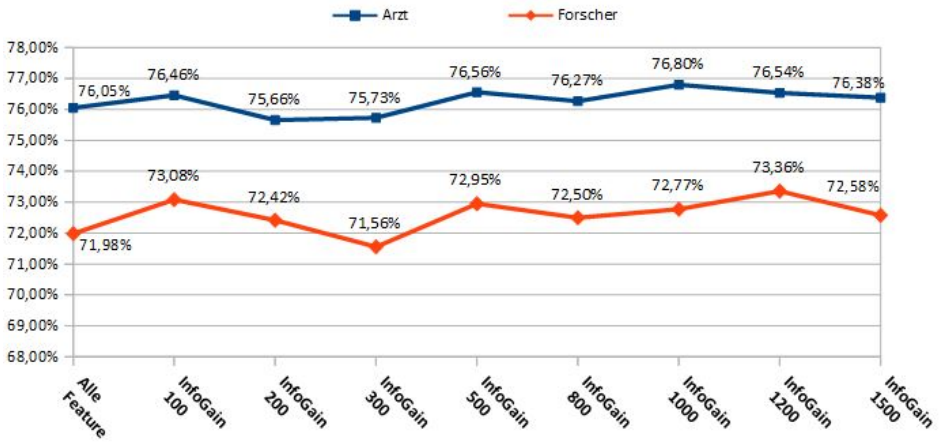


Abbildung 6.5: Vergleich verschiedener F1-Measure mit verschiedener Anzahl der Merkmale

6.3 Vergleich der Ergebnisse vor und nach Zusammenfassen der Klassen *Arzt* und *Forscher*

Im Abschnitt 6.2 wurden die zwei Klassen *Arzt* und *Forscher* zuerst zusammengefasst, anschließend wurde ein zweiter binärer Klassifikator entwickelt um die Klassen *Arzt* und *Forscher* zu erkennen. Das Histogramm 6.6 zeigt den Vergleich der Ergebnisse vor und nach dem Zusammenfassen der Klassen *Arzt* und *Forscher*, mit 10-Cross-Validation. Aus dem Vergleich wird ersichtlich, dass sich das Ergebnis der Klassen *Arzt* und *Forscher* 15% bis 20% verbessert hat. Das Ergebnis aller anderen Klassen (außer der Klasse *Other*) dagegen hat sich nicht wesentlich verändert. Es ist deswegen sinnvoll, zwei Klassifikatoren aufzubauen. Einer wird für die 7 Klassen (hier *AF, Behörden, Journalist, Patient, Pharma, Spam,* und *Other*) gebraucht und einer für die zwei Klassen *Arzt* und *Forscher*.

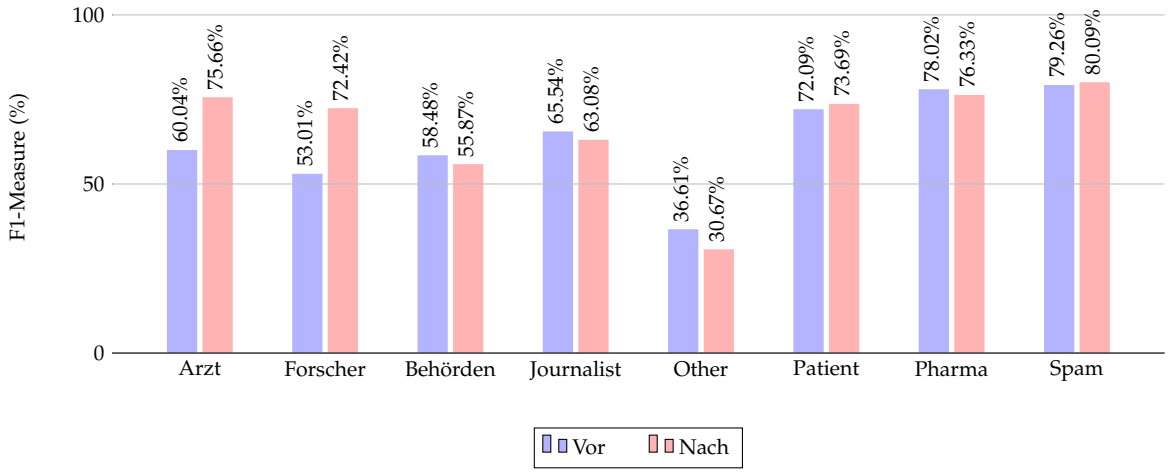


Abbildung 6.6: Vergleich vor und nach dem Verbinden der Klassen *Arzt* und *Forscher*

6.4 Evaluation durch Test-Set

Das Ziel der Evaluierung mit einer Testdatei bestand darin, die Präzision, Recall und F1-Measure der entwickelten Klassifikatoren zu überprüfen. 277 zufällig ausgewählte Tweets, die nicht in der Training-Set eingeschlossen sind, bildeten die Testdatei. Die zwei Klassifikatoren, die nach dem Zusammenfassen der Klassen *Arzt* und *Forscher* gebraucht wurden, wurden durch die Testdatei evaluiert. Die Abbildung 6.5 zeigt zwei Konfus-Matrix. Auf der linken Seite ist die Matrix des MaxEnt-Klassifikator, der mit der Methode *InfoGain200* entwickelt wurde. Mit ihm werden die 7 Klassen (nach dem Zusammenfassen der Klassen *Arzt* und *Forscher*) erkannt. Auf der rechten Seite ist die Matrix des binären naiven Bayes-Klassifikator, mit dem die Klassen *Arzt* und *Forscher* erkannt werden. Die Angaben in der Abbildung 6.6 beziehen sich auf die Präzision, Recall und F1-Measure jeder Klasse. Davon haben die Klassen *Behörden, Pharma* und *Spam* eine F1-Measure von über 90%. Die Klassen *Arzt* und *Forscher* haben jeweils ein F1-Measure von 88,8% und 83,6%.

label	AF	Behoerden	Journalist	Other	Patient	Pharma	Spam	total
AF	70	.	3	.	9	.	1	83
Behoerden	1	13	.	.	.	.	.	14
Journalist	13	.	30	.	.	3	3	49
Other	.	.	.	7	9	.	1	17
Patient	5	.	.	.	45	.	.	50
Pharma	.	.	.	.	.	20	.	20
Spam	1	.	1	.	.	.	42	44

label	Arzt	Forscher	total
Arzt	44	3	47
Forscher	9	27	36

Tabelle 6.5: links: Konfus-Matrix der 7 Klassen (*Arzt* und *Forscher* wurden zusammengefasst): Durchschnitt der Train-Genauigkeit = 98,78%, Durchschnitt der Test-Genauigkeit mean = 81,95%; **rechts:** Konfus-Matrix der Klassen *Arzt* und *Forscher*: Durchschnitt der Train-Genauigkeit = 96,10%, Durchschnitt der Test-Genauigkeit mean = 85,54%

	Präzision	Recall	F1
Arzt	83.0%	93.6%	88.0%
Forscher	90.0%	75.0%	81.8%
Behörden	100%	92.9%	96.3%
Journalist	88.2%	61.2%	72.2%
Other	100%	41.2%	58.3%
Patient	71.4%	90.0%	79.6%
Pharma	86.9%	100%	93.0%
Spam	89.3%	95.5%	92.3%

Tabelle 6.6: Die Präzision, Recall und F1-Measure der 8 Klassen.

Lernkurve des Corpus

Um abzuschätzen, wie viel Trainingsdaten ausreichend sind, wird hierfür eine Lernkurve des Corpus generiert. Die Corpus aus Kapitel 5 werden in fünf Teile gegliedert von denen jeder 530-540 Tweets enthält. Die Abbildung 6.7 zeigt, dass die Leistung des Systems am besten ist, wenn alle Trainingsdaten verwendet werden. Mit dem Wachstum der Trainingsdatei wird eine immer höhere Genauigkeit erreicht und der Trend der Steigerungsrate verlangsamt sich. Somit wird in dem aktuellen Zustand eine vergrößerte Menge Trainingsdaten der Leistung des Systems zu Verbesserung helfen.

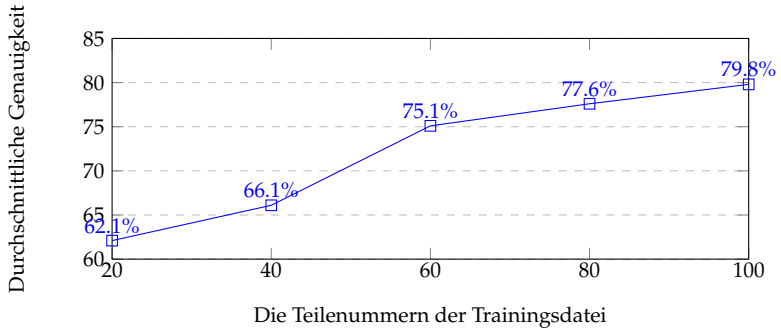


Abbildung 6.7: Lernkurve des Corpus

6.5 Fazit

Mehrere Klassifikatoren mit verschiedenen Methoden zur Auswahl der Merkmale wurden in diesem Kapitel entwickelt und durch 10-Cross-Validation evaluiert.

- **Methoden der Auswahl der Merkmalen** Die Methode N-Gram ist zwar einfach, aber sie berücksichtigt nur die Positionsbeziehungen der Wörter. Die Ähnlichkeit zwischen Wörter und Semantik ist hierin nicht eingeschlossen. Die Term-Gewichtung-Methoden Mutual Information (MI) und Information Gain (InfoGain) liefern in dieser Arbeit gute Ergebnisse. MI und InfoGain sind in gewissem Maße ähnlich, daher werden MI und InfoGain in dieser Arbeit kombiniert. Ihr Unterschied liegt darin, dass die Bedeutung von InfoGain sich nach erhöhten Informationen, nach der Klassifikation des Systems und MI nach der Information zwischen Term und Klasse richtet. Wegen der Spärlichkeit der kurzen Texte wurde das LDA-Topic-Model zusätzlich eingeführt. Das resultat zeigt, dass InfoGain& MI und LDA-Model die F1-Measure aller Klassen um 1% bis 3% verbessert haben.
- **Wirkungen der Klassifikatoren** Bei der Kategorisierung von Multi-Klassen erreicht der Klassifikator Maximum Entropy eine bessere Wirkung als Naiv Bayes. Außerdem schneidet MaxEnt besser als Naiv Bayes ab, wenn die Anzahl der Texte ungleich in verschiedenen Klassen verteilt sind. Im Fall der Klassifikation von binärer Klassen führt Naiv Bayes zu guten Ergebnissen.
- Zum Abschluss wurden die zwei relativ ähnliche Klassen *Arzt* und *Forscher* verbunden. Die F1-Measure der neuen Klasse hat sich im Vergleich zur Klasse *Arzt* oder *Forscher* um 12% bis 20% verbessert. Bei einer binären Klassifikation zwischen *Arzt* und *Forscher* liefert die F1-Measure jeweils 76% und 72%.

7 Zusammenfassung und Ausblick

7.1 Zusammenfassung

Das Ziel dieser Arbeit ist, die Tweets, die Informationen über populärere Arzneimittel enthalten, automatisch zu kategorisieren. Es wird dadurch ermittelt, welche sozialen Gruppen über welche Arzneimittel am meisten in Twitter gepostet haben. Dafür wurden verschiedene Klassifikatoren entwickelt, welche die verschiedenen Merkmale verwenden.

- Zuerst wurde eine Liste der meistverkauften und am häufigsten verschriebenen Arzneimittel bestimmt, die von dem medizinischen Magazin *Pharmacy Times* angeboten wurden. Mit Hilfe der Informationen von DRUGBANK wurde die Liste der Arzneimitteln durch die entsprechenden Synonyma ergänzt. Unter der Verwendung der Technik StAX und JSON wurden die Medikamentennamen aus der XML-Datei gelesen und danach in JSON-Form gespeichert.

- Anschließend wurde ein Java-Projekt mithilfe der Twitter Search API entwickelt, mit dem die Tweets von 2008 bis 2015 gesammelt werden können. Von den gesammelten Tweets wurden 2969 Tweets zufällig ausgewählt und mit entsprechender vordefinierter Klasse gekennzeichnet und als Corpus zur Klassifikation verwendet. Von den 2969 Tweets bilden 277 Tweets die Testdatei.

- Die Tweets wurden zuerst normiert, ohne ihre Inhalte zu ändern. Aus ihnen wurden die Merkmalen zur Klassifikation bestimmt. Mallet bietet einen, aus Information Gain und Mutual Information kombinierten Algorithmus an. Mit Information Gain wurden die Top 1000 Wörter als Merkmale bestimmt und anschließend mit Mutual Information sortiert. Die Top-100, Top-200, Top-300, Top-500, Top-800 und Top-1000 von den 1000 Wörtern wurden in sechs Durchläufen als Merkmale zur Kategorisierung benutzt. Nach dem Vergleich liefert der Klassifikator mit Top-200 Wörtern das beste Ergebnis. Ausgehend von der Spärlichkeit der kurzen Texte wurde die Mehtode LDA-Topic-Model auch zur Auswahl der Merkmale verwendet. Jede Klasse hat ihr eigenes Thema. Die Themen wurden in vier Durchläufen, mit jeweils einer verschiedenen Anzahl an Wörter, trainiert. Beim ersten mal bestehen die Themen jeder Klasse aus 100 Wörter, beim zweiten aus 200 Wörter, beim dritten aus 300 Wörter und beim vierten mal aus 500 Wörter. Hierbei stellte sich heraus, dass der Klassifikator, dessen Themen mit 200 Wörtern trainiert wurden, sich am besten zur Kategorisierung eignet. Die Kombination der Methoden InfoGain & MI und LDA-Topic-Model erweitern die Methoden der Auswahl der Merkmale. Eine Kombination, die aus den Top-100 Wörter der Methode InfoGain & MI und den Themen die mit 100 Wörter trainiert wurden besteht, funktioniert im Vergleich zu den anderen Kombinationen am besten .

Die Methode N-Gram (in dieser Arbeit nur 2-Gram und 3- Gram) wurden auch zur Auswahl der Merkmale benutzt, führten aber bei der F1-Measure zu keinen guten Ergebnissen. Daraus folgt, dass die Methode N-Gram bei der Kategorisierung kurzer Texte nicht gut anwendbar ist.

- Die zwei inhaltlich ähnlichen Klassen *Arzt* und *Forscher* wurden in eine neuen Klasse **„ AF“** zusammengefasst. Die F1-Measure der neuen Klassen **„ AF“** hat sich im Vergleich zu den separaten Klassen *Arzt* und *Forscher* um 12% bis 20% verbessert. Die F1-Measure einer binären Klassifikation der Klassen *Arzt* und *Forscher* führte auch zu einem guten Ergebnis von 76,27% für die Klasse *Arzt* und 72,58% für die Klasse *Forscher*.
- Für diese Arbeit wurde hauptsächlich der Maximum Entropy Klassifikator verwendet, da der MaxEnt-Klassifikator bei den Klassen, welche aus einer kleinen Menge von Tweets bestehen, besser als Naiv Bayes und C45 Desision Tree kategorisieren kann. Bei der binären Klassifikation der Klassen *Arzt* und *Forscher* wurde der Naiv Bayes Klassifikator benutzt. Im Vergleich zu anderen Klassifikatoren liefert er die besten Ergebnisse.
- Im Corpus dieser Arbeit wurden die Nutzer den Klassen *Arzt, Paient, Journalist und Spam* zugeordnet. Die vier Klassen haben zwischen einander relativ viele gemeinsamen Arzneimittel. Die am meisten diskutierten Arzneimittel im Corpus sind zur Behandlungen von Schmerzen, Krebs/Tumor, psychischer Störung, Antibiotikum, Diabetes und Herz-Kreislauf-Krankheiten sowie Allergien.

Zusammenfassend kann gesagt werden, dass mehrere Klassifikatoren mit den verschiedenen Methoden zur Auswahl der Merkmale in dieser Arbeit aufgebaut und ihre Ergebnisse der F1-Measure durch Cross-Validation verglichen wurden. Dabei wurde die Testdatei der Tweets aus allen Klassen, mit einer F1-Measure von über 50% korrekt erkannt.

7.2 Ausblick

- Die Ergebnisse der Klassifikatoren in der Evaluation zeigten, unabhängig davon welches Verfahren verwendet wurde, dass die Erkennung von Tweets der Klassen *Behörden* und *Other* noch nicht zufriedenstellend ist. Dies liegt für die Klasse *Behörden* hauptsächlich an der Anzahl der gesammelten Tweets, welche nur von den bestimmten Organisationen oder Instituten der nationalen/staatlichen Ebene gepostet wurden. Grund der schlechten Erkennung der Klasse *Other*, welche von allen unterschiedlichen Benutzeridentitäten gepostet werden, ist, dass die Tweets dieser Klasse inhaltlich sehr verworren und nicht explizit sind. Der Klassifikator könnte eine so genannte „ ausschließende Regel “ brauchen, mit der ein Tweet, wenn er keiner der 7 Klassen zugeordnet werden kann, der Klasse *Other* zugeordnet wird.
- Die Konfus-Matrizen in der Evaluation zeigten, dass es bei der Erkennung der Klassen *Arzt*, *Forscher* und *Patient* leicht zu Verwechselungen kommt. Es wäre beispielsweise sinnvoll bei der Sammlung auf Twitter die Profile der entsprechenden Benutzer des

Tweets auch zu sammeln, da viele Twitter Nutzer gern ihre Berufe und Beschäftigungen in ihrer Profile schreiben. Somit würde der Klassifikator bei der Klassifikation auch die Stichwörter im Profil berücksichtigen und könnte damit die Erkennung verbessern.

- Fast jeder Tweet der Klassen *Journalist* und *Spam* enthält zumindest eine URL. Wenn ein spezieller Analyser, basierend auf Java regulärem Ausdruck, dazu verwendet werden würde, die URLs in den gesammelten Tweets zu analysieren und zu erkennen, wäre das hilfreich.
- Es ist sinnvoll, das Corpus immer weiter zu erweitern.
- Durch 10-Cross-Validation scheint die Stabilität des MaxEnt-Klassifikators nicht so gut wie Naiv Bayes zu sein. Das Problem muss mit weiteren Untersuchungen und experimentellem Umfang überprüft werden.

8 Appendix

Liste der Arzneimittel

```
{"primyname":"Hydrocodone lang:en",
"synonym": [
"Dihydrocodeinone lang:en",
"Hidrocodona lang:en",
"Hydrocodon lang:en",
"Hydrocodonum lang:en",
"Hydrocone lang:en",
"Hydroconum lang:en",
"Idrocodone lang:en"]}
,
{"primyname":"Levothyroxine Sodium lang:en",
"synonym": [
"L-Thyroxin lang:en",
"L-Thyroxin Henning lang:en",
"L-T4 lang:en",
"L-Thyroxine lang:en",
"Levothyroxin lang:en",
"Levothroid lang:en",
"Levoxyl lang:en",
"Synthroid lang:en",
"Tirosint lang:en",
"Unithroid lang:en"]}
,
{"primyname":"Azithromycin lang:en",
"synonym": [
"Azenil lang:en",
"Azifast lang:en",
"Azigram lang:en",
"Azimakrol lang:en",
"Azithromycine lang:en",
"Azithromycinum lang:en",
"Azitromicina lang:en",
"Azitromin lang:en",
"Hemomycin lang:en",
"Zithromax lang:en",
"Zmax lang:en"]}
,
{"primyname":"Celebrex lang:en",
"synonym": [
"Celecoxib lang:en",
"Celecoxibum lang:en"]}
,
{"primyname":"Januvia lang:en",
"synonym": [
"Sitagliptan lang:en",
"Sitagliptin phosphate lang:en",
"Sitagliptina lang:en",
"Sitagliptine lang:en",
"Sitagliptinum lang:en"]}
,
{"primyname":"Amoxicillin lang:en",
"synonym": [
"Amolin lang:en",
"Amopenixin lang:en",
"Amox lang:en",
"Amoxicilina lang:en",
"Amoxicillin lang:en",
"Amoxicillin anhydrous lang:en",
"Amoxicilline lang:en",
"Amoxicillinum lang:en",
"Amoxycillin lang:en",
"Clamoxyl lang:en",
"Moxal lang:en",
"p-Hydroxyampicillin lang:en"]}
,
{"primyname":"Cymbalta lang:en",
"synonym": [
"Duloxetine lang:en",
"LY 248686 lang:en"]}
,
{"primyname":"Venlafaxine lang:en",
"synonym": [
"Elafax lang:en",
"Venlafaxina lang:en",
"Venlafaxinum lang:en"]}
,
{"primyname":"Sensipar lang:en",
"synonym": [
"Cinacalcet lang:en",
"Mimpara lang:en"]}
,
{"primyname":"Lovenox lang:en",
"synonym": [
"Enoxaparin lang:en",
"LMWH lang:en",
"Low Molecular Weight Heparin lang:en"]}
,
{"primyname":"Lyrica lang:en",
"synonym": [
"pregabalin lang:en"]}
,
{"primyname":"Namenda lang:en",
"synonym": [
"1-Amino-3,5-dimethyladamantane lang:en",
"Memantina lang:en",
"Memantine lang:en",
"Memantinum lang:en",
"Memantine lang:en"]}
,
{"primyname":"Vyvanse lang:en",
"synonym": [
"Lisdexamfetamine lang:en",
"lisdexamfetamine dimesylate lang:en",
"NRP104 lang:en"]}
,
{"primyname":"Synthroid lang:en",
"synonym": [
"Levothyroxine lang:en",
"3,3',5,5'-Tetraiodo-L-thyronine lang:en",
"L-T4 lang:en",
"L-Thyroxine lang:en",
"Levothyroxin lang:en",
"LT4 lang:en",
"T4 lang:en"]}
,
{"primyname":"Levaquin lang:en",
"synonym": [
"Levofloxacin lang:en",
"Ofloxacin lang:en",
"Levofloxacine lang:en",
"Levofloxacino lang:en",
"Levofloxacinum lang:en",
"Ofloxacin S-(-)-form lang:en",
"Rimantadin lang:en",
"Rimantadine lang:en"]}
,
{"primyname":"Lantus lang:en",
"synonym": [
"Lantus Solostar lang:en",
"Insulin Glargine (rDNA origin) lang:en",
"Insulin Glargine lang:en"]}
,
{"primyname":"Omeprazole lang:en",
"synonym": [
"Prilosec lang:en"]}
,
{"primyname":"Gleevec lang:en",
"synonym": [
"Imatinib lang:en",
"Imatinibum lang:en",
"STI 571 lang:en"]}
,
{"primyname":"Abilify lang:en",
"synonym": [
"Aripiprazole lang:en",
"Abilitat lang:en",
"Aripiprazol lang:en",
"Aripiprazolum lang:en"]}
```

```
,
{"primyname":"Clonazepam lang:en",
"synonym": [
"Clonazepamum lang:en"]}
,
{"primyname":"Hydrochlorothiazide lang:en",
"synonym": [
"Esidrix lang:en",
"HCTZ lang:en",
"Microzide lang:en"]}
,
{"primyname":"Neupogen lang:en",
"synonym": [
"Filgrastim lang:en",
"G-CSF lang:en",
"Granulocyte Colony Stimulating Factor lang:en",
"Tbo-filgrastim lang:en"]}
,
{"primyname":"Lipitor lang:en",
"synonym": [
"Lipovastatinklonal lang:en",
"Atorvastatin lang:en"]}
,
{"primyname":"Plavix lang:en",
"synonym": [
"Clopidogrel lang:en",
"Clopidogrelum lang:en"]}
,
{"primyname":"Avastin lang:en",
"synonym": [
"Bevacizumab lang:en",
"antiVEGF lang:en"]}
,
{"primyname":"Xeloda lang:en",
"synonym": [
"Capecitabine lang:en",
"Capecitabin lang:en",
"Capecitabina lang:en",
"Capecitabinum lang:en"]}
,
{"primyname":"Remicade lang:en",
"synonym": [
"Infliximab lang:en"]}
,
{"primyname":"Detrol LA lang:en",
"synonym": [
"Tolterodina lang:en",
"Tolterodine lang:en",
"Tolterodinum lang:en"]}
,
{"primyname":"Betaseron lang:en",
"synonym": [
"Interferon beta-1b lang:en",
"Fibroblast interferon lang:en",
"IFN-beta lang:en",
"Interferon beta precursor lang:en"]}
,
{"primyname":"Aranesp lang:en",
"synonym": [
"Darbepoetin alfa lang:en",
"Epoetin lang:en",
"Erythropoietin precursor lang:en"]}
,
{"primyname":"Tarceva lang:en",
"synonym": [
"Erlotinib lang:en",
"OSI-774 lang:en",
"Gefitinib lang:en",
"ZD 1839 lang:en"]}
,
{"primyname":"Boniva lang:en",
"synonym": [
"Ibandronate lang:en",
"Bondronat lang:en",
"Ibandronic Acid lang:en"]}
,
{"primyname":"Nexium lang:en",
"synonym": [
"Esomeprazole lang:en",
"omeprazole lang:en",
"Esomeprazol lang:en",
"Esomeprazolum lang:en",
"Perprazole lang:en"]}
,
{"primyname":"Avonex lang:en",
"synonym": [
"Interferon beta-1a lang:en",
"IFN-beta lang:en",
"Interferon beta precursor lang:en"]}
,
{"primyname":"Evista lang:en",
"synonym": [
"Raloxifene lang:en",
"Keoxifene lang:en",
"Raloxifeno lang:en",
"Raloxifenum lang:en"]}
,
{"primyname":"Singulair lang:en",
"synonym": [
"Montelukast lang:en",
"Montelukastum lang:en"]}
,
{"primyname":"Epogen lang:en",
"synonym": [
"Epoetin alfa lang:en",
"Procrit lang:en"]}
,
{"primyname":"Gabapentin lang:en",
"synonym": [
"1-(Aminomethyl)cyclohexaneacetic acid lang:en",
"Gabapentin GR lang:en",
"Gabapentina lang:en",
"Gabapentine lang:en",
"Gabapentino lang:en",
"Gabapentinum lang:en",
"Gabapetin lang:en",
"Neurontin lang:en"]}
,
{"primyname":"Zometa lang:en",
"synonym": [
"Zoledronate lang:en",
"Reclast lang:en",
"Zoledronic acid lang:en"]}
,
{"primyname":"Actonel lang:en",
"synonym": [
"Acide ris dronique lang:en",
"Acido risedronico lang:en",
"Acidum risedronicum lang:en",
"Ridron lang:en",
"Risedronate lang:en",
"Risedronic acid lang:en",
"Risedrons ure lang:en"]}
,
{"primyname":"Cephalexin lang:en",
"synonym": [
"Cefalexin lang:en",
"Cefalexina lang:en",
"C falexine lang:en",
"Cefalexinum lang:en",
"Celexin lang:en",
"Cepastar lang:en",
"Cepexin lang:en",
"Cephacillin lang:en",
"C phalexine lang:en",
"Ceporexin lang:en"]}
,
{"primyname":"Zetia lang:en",
"synonym": [
"Ezetimibe lang:en",
"Ezedoc lang:en",
"Ezetimiba lang:en",
"Ezetimibum lang:en",
"Ezetrol lang:en"]}
,
{"primyname":"Simvastatin lang:en",
"synonym": [
"Simvastatina lang:en",
"Simvastatine lang:en",
"Simvastatinum lang:en",
"Synvinolin lang:en",
"Zocor lang:en"]}
,
{"primyname":"Atenolol lang:en",
"synonym": [
"Atenololum lang:en"]}
,
{"primyname":"Lexapro lang:en",
"synonym": [
"Escitalopram lang:en",
"Citalopram lang:en",
"Escitalopram Oxalate lang:en",
```

```
"Escitalopramum lang:en",
"Esertia lang:en"]}
,
{"primyname":"Prezista lang:en",
"synonym": [
"Darunavir lang:en",
"Darunavirum lang:en",
"TMC114 lang:en"]}
,
{"primyname":"Oxycontin lang:en",
"synonym": [
"Oxycodone lang:en",
"Oxicodona lang:en",
"Oxycodonum lang:en",
"14-Hydroxydihydrocodeinone lang:en",
"Dihydrohydroxycodeinone lang:en",
"Dihydroxycodeinone lang:en"]}
,
{"primyname":"Benicar lang:en",
"synonym": [
"Olmesartan lang:en",
"Olmesartan medoxomil lang:en"]}
,
{"primyname":"Levemir lang:en",
"synonym": [
"Insulin detemir lang:en",
"Insulin detemir recombinant lang:en",
"Levemir flexpen lang:en",
"Levemir innolet lang:en",
"Levemir penfill lang:en"]}
,
{"primyname":"Lucentis lang:en",
"synonym": [
"Ranibizumab lang:en",
"rhuFab V2 lang:en"]}
,
{"primyname":"Lunesta lang:en",
"synonym": [
"Eszopiclone lang:en",
"Zopiclone lang:en",
"Esopiclone lang:en",
"Estorra lang:en"]}
,
{"primyname":"Herceptin lang:en",
"synonym": [
"Trastuzumab lang:en",
"Anti HER2 lang:en"]}
,
{"primyname":"Humira lang:en",
"synonym": [
"Adalimumab lang:en"]}
,
{"primyname":"Diovan lang:en",
"synonym": [
"Valsartan lang:en"]}
,
{"primyname":"Carvedilol lang:en",
"synonym": [
"Carvedilolum lang:en"]}
,
{"primyname":"Reyataz lang:en",
"synonym": [
"Atazanavir lang:en",
"Atazanavirum lang:en",
"Zrivada lang:en"]}
,
{"primyname":"Cubicin lang:en",
"synonym": [
"Daptomycin lang:en"]}
,
{"primyname":"Rituxan lang:en",
"synonym": [
"Rituximab lang:en",
"AntiCD20 lang:en"]}
,
{"primyname":"Alimta lang:en",
"synonym": [
"Pemetrexed lang:en"]}
,
{"primyname":"Neulasta lang:en",
"synonym": [
"Pegfilgrastim lang:en",
"Lenograstim lang:en",
"Pluripoietin lang:en"]}
,
{"primyname":"Velcade lang:en",
"synonym": [
"Bortezomib lang:en"]}
,
{"primyname":"Furosemide lang:en",
"synonym": [
"Furosemid lang:en",
"Frusemide lang:en",
"Furosemida lang:en",
"Furosemidu lang:en",
"Furosemidum lang:en",
"Lasix (tn) lang:en"]}
,
{"primyname":"Copaxone lang:en",
"synonym": [
"Glatiramer Acetate lang:en"]}
,
{"primyname":"Spiriva HandiHaler lang:en",
"synonym": [
"Tiotropium lang:en"]}
,
{"primyname":"Meloxicam lang:en",
"synonym": [
"M loxicam lang:en",
"Meloxicamum lang:en",
"Mobic lang:en"]}
,
{"primyname":"Provigil lang:en",
"synonym": [
"Modafinil lang:en",
"Modafinilo lang:en",
"Modafinilum lang:en",
"Moderateafinil lang:en"]}
,
{"primyname":"Metoprolol Tartrate lang:en",
"synonym": [
"Metoprolol lang:en",
"Metoprolol Succinate lang:en"]}
,
{"primyname":"ProAir HFA lang:en",
"synonym": [
"Salbutamol lang:en",
"Levalbuterol lang:en",
"Proventil lang:en",
"albuterol inhalation lang:en",
"Albuterol lang:en",
"ProAir RespiClick lang:en",
"Proventil HFA lang:en",
"Ventolin HFA lang:en",]}
,
{"primyname":"Trazodone HCl lang:en",
"synonym": [
"Trazodona lang:en",
"Trazodone lang:en",
"Trazodonum lang:en",
"Oleptro lang:en",
"Desyrel lang:en",
"Desyrel Dividose lang:en"]}
,
{"primyname":"Advair Diskus lang:en",
"synonym": [
"Advair HFA lang:en",
"fluticasone and salmeterol lang:en"]}
,
{"primyname":"Tramadol HCl lang:en",
"synonym": [
"Tramadol lang:en",
"ConZip lang:en",
"Rybix ODT lang:en"]}
,
{"primyname":"Ibuprofen lang:en",
"synonym": [
"P-Isobutylhydratropic acid lang:en",
"2-(4-Isobutylphenyl)propanoic acid lang:en",
"4-Isobutylhydratropic acid lang:en",
"Amibufen lang:en",
"Anflagen lang:en",
"Apsifen lang:en",
"Brufen lang:en",
"Brufort lang:en",
"Buburone lang:en",
"Butylenin lang:en",
"Dolgirid lang:en",
"Dolgit lang:en",
"Dolo-dolgit lang:en",
"Dolormin lang:en",
"Ebufac lang:en",
```

```
"Epobron lang:en",
"Femadon lang:en",
"Ibu-attritin lang:en",
"IbuHexal 400 lang:en",
"Ibumetin lang:en",
"Ibuprocin lang:en",
"Ibuprophen lang:en",
"Lebrufen lang:en",
"Medipren lang:en",
"Motrin lang:en",
"Mynosedin lang:en",
"Naproxen lang:en",
"Nobfen lang:en",
"Nobgen lang:en",
"Nuprin lang:en",
"Nurofen lang:en",
"Pediaprofen lang:en",
"Roidenin lang:en",
"Seclodin lang:en"]}
,
{"primyname":"Amlodipine Besylate lang:en",
"synonym": [
"Amlodipine lang:en",
"Amlodipine free base lang:en",
"Amlodipino lang:en",
"Amlodipinum lang:en"]}
,
{"primyname":"Fluticasone Propionate lang:en",
"synonym": [
"Cutivate lang:en",
"Fluticason lang:en",
"Fluticasona lang:en",
"Fluticasone lang:en",
"Fluticasonum lang:en"]}
,
{"primyname":"Warfarin Sodium lang:en",
"synonym": [
"Warfarin lang:en",
"Coumafene lang:en",
"Zoocoumarin lang:en",
"Coumadin lang:en",
"Jantoven lang:en"]}
,
{"primyname":"Sertraline HCl lang:en",
"synonym": [
"Sertraline lang:en",
"Sertralina lang:en",
"Sertralinum lang:en"]}
,
{"primyname":"Pravastatin Sodium lang:en",
"synonym": [
"Pravastatin lang:en",
"Pravastatin acid lang:en",
"Pravastatina lang:en",
"Pravastatine lang:en",
"Pravastatinum lang:en"]}
,
{"primyname":"Clavulanate Potassium lang:en",
"synonym": [
"Clavulanate lang:en",
"Acide clavulanique lang:en",
"Acido clavulanico lang:en",
"Acidum clavulanicum lang:en",
"Clavulanic Acid lang:en",
"Clavulansaeure lang:en"]}
,
{"primyname":"Seroquel lang:en",
"synonym": [
"Quetiapine lang:en",
"Quetiapina lang:en",
"Quetiapine fumarate lang:en",
"Quetiapine hemifumarate lang:en",
"Quetiapinum lang:en"]}
,
{"primyname":"Zolpidem Tartrate lang:en",
"synonym": [
"Zolpidem lang:en",
"Zolpidemum lang:en"]}
,
{"primyname":"Metformin HCl lang:en",
"synonym": [
"Metformin lang:en",
"1,1-Dimethylbiguanide lang:en",
"Haurymellin lang:en",
"Metformina lang:en",
"Metformine lang:en",
"Metformine pamoate lang:en",
"Metforminum lang:en",
"N,N-Dimethylimidodicarbonimidic diamide lang:en"]}
,
{"primyname":"Prednisone lang:en",
"synonym": [
"1,2-Dehydrocortisone lang:en",
"Dehydrocortisone lang:en",
"Prednisona lang:en",
"Prednisonum lang:en"]}
,
{"primyname":"Fluconazole lang:en",
"synonym": [
"Biozole lang:en",
"Diflucan lang:en",
"Elazor lang:en",
"Fluconazol lang:en",
"Fluconazolum lang:en",
"Triflucan lang:en"]}
,
{"primyname":"ACTOS lang:en",
"synonym": [
"Pioglitazone lang:en",
"Pioglitazona lang:en",
"Pioglitazonum lang:en"]}
,
{"primyname":"Vitamin D lang:en",
"synonym": [
"sunshine vitamin lang:en"]}
,
{"primyname":"Fluoxetine HCl lang:en",
"synonym": [
"Fluoxetin lang:en",
"Fluoxetina lang:en",
"Prozac lang:en"]}
,
{"primyname":"Citalopram HBr lang:en",
"synonym": [
"Citalopram lang:en",
"Citadur lang:en",
"Nitalapram lang:en"]}
,
{"primyname":"Ciprofloxacin HCl lang:en",
"synonym": [
"Ciprofloxacin lang:en",
"Ciprofloxacine lang:en",
"Ciprofloxacino lang:en",
"Ciprofloxacinum lang:en"]}
,
{"primyname":"Lorazepam lang:en",
"synonym": [
"Ativan lang:en",
"Lormetazepam lang:en"]}
,
{"primyname":"Nasonex lang:en",
"synonym": [
"mometasone nasal lang:en",
"Mometason lang:en",
"Mometasona lang:en",
"Mometasone Furoate lang:en",
"Mometasone Furoate Hydrate lang:en",
"Mometasonfuroat lang:en",
"Mometasoni furoas lang:en",
"Mometasonum lang:en",
"Mom tasone lang:en",
"Mom tasone (furoate de) lang:en"]}
,
{"primyname":"Tricor lang:en",
"synonym": [
"Fenofibrate lang:en",
"Fenofibrato lang:en",
"Fenofibratum lang:en",
"Fenofibric acid lang:en",
"Finofibrate lang:en",
"Lipantil (tn) lang:en"]}
,
{"primyname":"Atripla lang:en",
"synonym": [
"Efavirenz lang:en",
"Efavirenzum lang:en"]}
,
{"primyname":"Truvada lang:en",
"synonym": [
"Emtricitabine lang:en",
"Emtricitabin lang:en",
"Emtricitabina lang:en",
```

```
"Emtricitabinum lang:en"]}
,
{"primyname":"Enoxaparin Sodium lang:en",
"synonym": [
"Enoxaparin lang:en"]}
,
{"primyname":"Geodon lang:en",
"synonym": [
"Ziprasidone lang:en",
"Ziprasidona lang:en",
"Ziprasidonum lang:en"]}
,
{"primyname":"Suboxone lang:en",
"synonym": [
"buprenorphine and naloxone lang:en",
"Naloxone lang:en",
"L-Naloxone lang:en",
"Naloxona lang:en",
"Naloxonum lang:en",
"Buprenorphine lang:en",
"Buprenorfina lang:en",
"Buprenorphinum lang:en"]}
,
{"primyname":"Lidoderm lang:en",
"synonym": [
"Lidocaine lang:en",
"Lignocaine lang:en"]}
,
{"primyname":"Eloxatin lang:en",
"synonym": [
"Oxaliplatin lang:en",
"Oxalatoplatinum lang:en"]}
,
{"primyname":"Niaspan lang:en",
"synonym": [
"Niacin lang:en",
"3-carboxypyridine lang:en",
"3-Pyridinecarboxylic acid lang:en",
"Acide Nicotinique lang:en",
"Acido nicotinico lang:en",
"Acidum Nicotinicum lang:en",
"Anti-pellagra vitamin lang:en"]}
,
{"primyname":"Androgel lang:en",
"synonym": [
"Testosterone lang:en",
"Androderm lang:en",
"Depo-Testadiol lang:en",
"Mertestate lang:en",
"Synandrol F lang:en",
"Testosteron lang:en",
"Testosterona lang:en",
"Testosteronum lang:en",
"Testost rone lang:en",
"Testoxyl lang:en",
"Testryl lang:en",
"Virosterone lang:en"]}
,
{"primyname":"Combivent lang:en",
"synonym": [
"albuterol and ipratropium lang:en",
"Ipratropium bromide lang:en"]}
,
{"primyname":"Rebif lang:en",
"synonym": [
"Interferon beta-1a lang:en",
"Interferon beta precursor lang:en"]}
,
{"primyname":"Symbicort lang:en",
"synonym": [
"budesonide and formoterol lang:en"]}
,
{"primyname":"NovoLog lang:en",
"synonym": [
"NovoLOG PenFill lang:en",
"NovoLOG FlexTouch lang:en",
"NovoLog Flexpen lang:en",
"Insulin Aspart lang:en",
"Aspart lang:en",
"Aspart Insulin lang:en",
"Insulin X14 lang:en",
"Insulin, Asp(B28) lang:en"]}
,
{"primyname":"Lovaza lang:en",
"synonym": [
"omega-3 polyunsaturated fatty acids lang:en",
"Krill Oil lang:en"]}
,
{"primyname":"Humalog lang:en",
"synonym": [
"Insulin Lispro lang:en",
"Insulin Lispro Recombinant lang:en"]}
,
{"primyname":"Adderall XR lang:en",
"synonym": [
"Adderall lang:en",
"amphetamine and dextroamphetamine lang:en"]}
,
{"primyname":"Aciphex lang:en",
"synonym": [
"Clofezone lang:en",
"Rabeprazole lang:en"]}
,

{"primyname":"Concerta lang:en",
"synonym": [
"Daytrana lang:en",
"Methyl phenidylacetate lang:en",
"Methylphenidan lang:en",
"Methylphenidatum lang:en",
"Metilfenidato lang:en"]}
,
{"primyname":"Budesonide lang:en",
"synonym": [
"Entocort EC lang:en",
"Uceris lang:en"]}
,
{"primyname":"Incivek lang:en",
"synonym": [
"telaprevir lang:en"]}
,
{"primyname":"Varivax lang:en",
"synonym": [
"varicella virus lang:en"]}
,
{"primyname":"Prevnar 13 lang:en",
"synonym": [
"pneumococcal 13-valent conjugate vaccine lang:en"]}
,
{"primyname":"Solodyn lang:en",
"synonym": [
"Minocycline lang:en",
"Minociclina lang:en",
"Minociclinum lang:en",
"Minocyclin lang:en",
"Minocyclinum lang:en",
"Dynacin lang:en",
"Minocin lang:en",
"Minocin PAC lang:en",
"Vectrin lang:en",
"Myrac lang:en"]}
,
{"primyname":"Isentress lang:en",
"synonym": [
"Raltegravir lang:en"]}
,
{"primyname":"Janumet lang:en",
"synonym": [
"metformin and sitagliptin lang:en",
"Sitagliptin lang:en",
"Janumet XR lang:en"]}
,
{"primyname":"Methylphenidate ER lang:en",
"synonym": [
"Methylphenidate lang:en",
"alpha-Phenyl-2-piperidineacetic acid methyl ester lang:
    en",
"Daytrana lang:en",
"Methyl phenidylacetate lang:en",
"Methylphenidan lang:en",
"Methylphenidatum lang:en",
"Metilfenidato lang:en",
]}
,
{"primyname":"Synagis lang:en",
"synonym": [
"Palivizumab lang:en"]}
,
{"primyname":"Restasis lang:en",
"synonym": [
"Cyclosporine lang:en",
"Ciclosporin lang:en",
```

```
"Cyclosporin A lang:en",
"cyclosporine ophthalmic lang:en"]}
,
{"primyname":"Erbitux lang:en",
"synonym": [
"Cetuximab lang:en",
"Anti EGFR lang:en"]}
,
{"primyname":"Vesicare lang:en",
"synonym": [
"Solifenacin lang:en",
"solifenacin succinate lang:en"]}
,
{"primyname":"Opana ER lang:en",
"synonym": [
"Oxymorphone lang:en",
"14-Hydroxydihydromorphinone lang:en",
"Dihydrohydroxymorphinone lang:en",
"Dihydroxymorphinone lang:en",
"Numorphan lang:en",
"Opana lang:en",
"Oximorphonum lang:en",
"Oxymorphine lang:en"]}
,
{"primyname":"Orencia lang:en",
"synonym": [
"Abatacept lang:en"]}
,
{"primyname":"Enbrel lang:en",
"synonym": [
"Etanercept lang:en",
"CD120b lang:en",
"TNF-R2 lang:en",
"Tumor necrosis factor receptor 2 lang:en",
"Tumor necrosis factor receptor superfamily member 1B
     precursor lang:en",
"Tumor necrosis factor receptor type II lang:en"]}
,
{"primyname":"Alprazolam lang:en",
"synonym": [
"8-Chloro-1-methyl-6-phenyl-4H-S-triazolo(4,3-a)(1,4)
     benzodiazepine lang:en",
"Xanax lang:en"]}
,
{"primyname":"Lisinopril lang:en",
"synonym": [
"Claritin lang:en",
"[N2-[(S)-1-CARBOXY-3-phenylpropyl]-L-lysyl-L-proline
     lang:en",
"Lisinopril anhydrous lang:en",
"Loratadina lang:en",
"Loratadinum lang:en"]}
,
{"primyname":"Zyprexa lang:en",
"synonym": [
"Olanzapine lang:en",
"Olanzapin lang:en",
"Olanzapina lang:en",
"Olanzapinum lang:en",
"2-methyl-4-(4-methyl-1-piperazinyl)-10H-thieno[2,3-b
     ][1,5] lang:en"]}
,
{"primyname":"Zyvox lang:en",
"synonym": [
"Linezolid lang:en",
"Linezolide lang:en",
"Linezolidum lang:en",
"N-(((S)-3-(3-Fluoro-4-morpholinophenyl)-2-oxo-5-
     oxazolidinyl)methyl)acetamide lang:en"]}
,
{"primyname":"Diazepam lang:en",
"synonym": [
"7-chloro-1,3-dihydro-1-methyl-5-phenyl-2H-1,4-
     benzodiazepin-2-one lang:en",
"Methyl diazepinone lang:en",
"Valium lang:en"]}
,
{"primyname":"Sulfamethoxazole and Trimethoprim lang:en
     ",
"synonym": [
"Sulfamethoxazole lang:en",
"Gantanol (tn) lang:en"]}
```

Literaturverzeichnis

[AAD14] J. W. Ayers, B. M. Althouse, and M. Dredze. Could behavioral medicine lead the web data revolution? *JAMA*, 311(14):1399–1400, 2014.

[Ble] D. M. Blei. Introduction to Probabilistic Topic Models.

[BNJ03] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[BPP96] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[BYRN+99] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[CCK14] A. A. Ciociola, L. B. Cohen, and P. Kulkarni. How drugs are developed and approved by the FDA: current process and future directions. *American Journal of Gastroenterology*, 109(5), 2014.

[CHWE12] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll. Sparse discriminant analysis. *Technometrics*, 2012.

[CT+94] W. B. Cavnar, J. M. Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

[CT12] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[DLY+07] L.-y. Dong, G.-y. Liu, S.-m. Yuan, P. Z. Li, et al. Classifier learning algorithm based on genetic algorithms. In *Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on*, pages 126–126. IEEE, 2007.

[FD13] S. Fox and M. Duggan. One in three American adults have gone online to figure out a medical condition. *Pew Internet & American Life Project*, 2013.

[Gam04] M. Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, page 841. Association for Computational Linguistics, 2004.

[GO06] R. Gutierrez-Osuna. Lecture 13: Validation. *Retrieved February*, 28:2007, 2006.

[K+95] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

[KKO+15] S. Kanouchi, M. Komachi, N. Okazaki, E. Aramaki, and H. Ishikawa. Who caught a cold?-identifying the subject of a symptom. In *Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics*, 2015.

[KLJ+11] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041, 2011.

[LHPD12] C. Lin, Y. He, C. Pedrinaci, and J. Domingue. Feature lda: a supervised topic model for automatic detection of web api documentations from the web. In *The Semantic Web–ISWC 2012*, pages 328–343. Springer, 2012.

[LKD+14] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2014.

[LWS+10] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards Internet-age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-related Social Networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 117–125, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[MC11] M. Mccord and M. Chuah. Spam detection on twitter using traditional classifiers. In *Autonomic and trusted computing*, pages 175–186. Springer, 2011.

[Moo01] A. W. Moore. Cross-validation for detecting and preventing overfitting. *School of Computer Science Carneigie Mellon University*, 2001.

[MRS+08] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[Nak13] G. Nakhaeizadeh. *Data Mining: Theoretische Aspekte und Anwendungen*, volume 27. Springer-Verlag, 2013.

[PD11] M. J. Paul and M. Dredze. You are what you Tweet: Analyzing Twitter for public health. *ICWSM*, 20:265–272, 2011.

[Pow11] D. M. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011.

[PP10] A. Pak and P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC*, volume 10, pages 1320–1326, 2010.

[RJX+05] L. Ronglu, W. Jianhui, C. Xiaoyun, T. Xiaopeng, and H. Yunfa. Using maximum entropy model for Chinese text categorization [J]. *Journal of Computer Research and Development*, 1:22–29, 2005.

[SGM] S. W. Shawn Graham and I. Milligan. Getting Started with Topic Modeling and MALLET.

[SLC12] H. Sampathkumar, B. Luo, and X.-w. Chen. Mining adverse drug side-effects from online medical forums. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*, pages 150–150. IEEE, 2012.

[TSK+06] P.-N. Tan, M. Steinbach, V. Kumar, et al. *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston, 2006.

[Wal] S. Waltl. Der Satz von Bayes.

[WKG+08] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906, 2008.

[Yan99] Y. Yang. An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69–90, 1999.

[YL99] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 1999.

All links were last followed on 9. Mai 2016

Declaration

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Stuttgart, 9. Mai 2016 _____

(Min Xu)