



**FOM Hochschule für Oekonomie & Management**

Hochschulzentrum Münster

## **Hausarbeit**

im Studiengang Big Data & Business Analytics

im Rahmen der Lehrveranstaltung

**Analyse semi- & unstrukturierter Daten**

über das Thema

### **CAPTUM**

**- Characterisation of Type IIb autoimmune chronic spontaneous urticaria markers -**

von

Fiete Ostkamp, Tim Lapstich und Artur Gergert

Betreuer : Prof. Dr. Rüdiger Buchkrämer

Matrikelnummern : 557851, , 562394

Abgabedatum : 4. August 2021

# Inhaltsverzeichnis

<b>Abbildungsverzeichnis</b>	<b>III</b>
<b>Tabellenverzeichnis</b>	<b>IV</b>
<b>Abkürzungsverzeichnis</b>	<b>V</b>
<b>Symbolverzeichnis</b>	<b>VI</b>
<b>1 Einleitung</b>	<b>1</b>
1.1 Zielsetzung . . . . .	1
1.2 Aufbau der Arbeit . . . . .	1
<b>2 Grundlagen der Auswertung unstrukturierter Daten</b>	<b>3</b>
2.1 Text Preprocessing . . . . .	3
2.2 Named Entity Recognition . . . . .	6
2.3 Named Entity Linking . . . . .	6
<b>3 Praxis</b>	<b>7</b>
3.1 Markeranalyse . . . . .	7
3.2 Überführung in Tabellenstruktur . . . . .	7
3.3 Markerkorrelationen . . . . .	7
3.4 Aufstellung des Gratingsystems . . . . .	7
<b>4 Fazit</b>	<b>7</b>
<b>Anhang</b>	<b>8</b>
<b>Literaturverzeichnis</b>	<b>9</b>

## Abbildungsverzeichnis

1	Verzeichnisstruktur der $\text{\LaTeX}$ -Dateien . . . . .	2
2	Text Mining Prozess . . . . .	3
3	Stemming vs. Lemmatization . . . . .	6

## **Tabellenverzeichnis**

## **Abkürzungsverzeichnis**

<b>NLP</b>	Natural Language Processing
<b>KDT</b>	“Knowledge Discovery in Textual Databases“

## **Symbolverzeichnis**

# 1 Einleitung

## 1.1 Zielsetzung

Die chronische spontane Urtikaria gehört zu der Gruppe chronischer Urtikaria Erkrankungen. Gekennzeichnet ist diese durch das Wiederauftreten von Quaddeln und/oder Angioödemem über einen Zeithorizont von mehr als sechs Wochen<sup>1</sup>. Die geschätzte weltweite Prävalenz chronischer Urtikaria Erkrankten beträgt schätzungsweise 1%. Es liegt lediglich eine geschätzte Prävalenz vor, da es Schwierigkeiten bei der Klassifizierung, der Identifizierung sowie der Diagnose der Erkrankung gibt. Dies ist vor allem auf erhebliche Verzögerungen bei der Diagnose sowie unzureichende Kenntnisse über die chronische Urtikaria zurückzuführen<sup>2</sup>.

Die oben angeführte Problematiken wurden zum Anlass genommen ein Projekt zu initiieren, welches den Auftrag verfolgt einen Beitrag zur bekämpfung der chronischen spontanen Urticaria Krankheit zu leisten. Begleitet wird das Projekt von Ärzten und Spezialisten der Charité in Berlin.

Die vorliegende Hausarbeit behandelt das Teilprojekt „Information Retrieval“. Ziel dieses Teilprojektes war es aus einem Text-Corpus mit insgesamt über 500 medizinische Fachartikel automatisiert Informationen aus den Texten zu extrahieren, um das Wissen über die Krankheit, erfolgreiche Behandlungsmöglichkeiten etc. zu erweitern.

## 1.2 Aufbau der Arbeit

- Grundlagen der Textvorverarbeitung - Lemmatisieren - Tokenisation - Schlüsselwörter extrahieren - Eigenes Sprachmodel auf Urticaria-Texte anwenden.







Kapitel 2 enthält die Inhalte des Thesis-Days und alles, was zum inhaltlichen erstellen der Thesis relevant sein könnte. In Kapitel 3 Praxis findet ihr wichtige Anmerkungen zu  $\LaTeX$ , wobei die wirklich wichtigen Dinge im Quelltext dieses Dokumentes stehen (siehe auch die Verzeichnisstruktur in Abbildung 1).

---

<sup>1</sup> Vgl. *Savic, S. et al.*, 2020, S. 4.

<sup>2</sup> Vgl. *ebd.*, S. 4.

**Abbildung 1: Verzeichnisstruktur der  $\text{\LaTeX}$ -Dateien**

Name	Änderungsdatum	Typ	Größe
 abbildungen	29.08.2013 01:25	Dateiordner	
 kapitel	29.08.2013 00:55	Dateiordner	
 literatur	31.08.2013 18:17	Dateiordner	
 skripte	01.09.2013 00:10	Dateiordner	
 compile.bat	31.08.2013 20:11	Windows-Batchda...	1 KB
 thesis_main.tex	01.09.2013 00:25	LaTeX Document	5 KB

Quelle: Eigene Darstellung



## 2 Grundlagen der Auswertung unstrukturierter Daten

Siehe auch Wissenschaftliches Arbeiten<sup>3</sup>. Damit sollten alle wichtigen Informationen abgedeckt sein ;-)<sup>4</sup> Hier gibt es noch ein Beispiel für ein direktes Zitat<sup>5</sup>

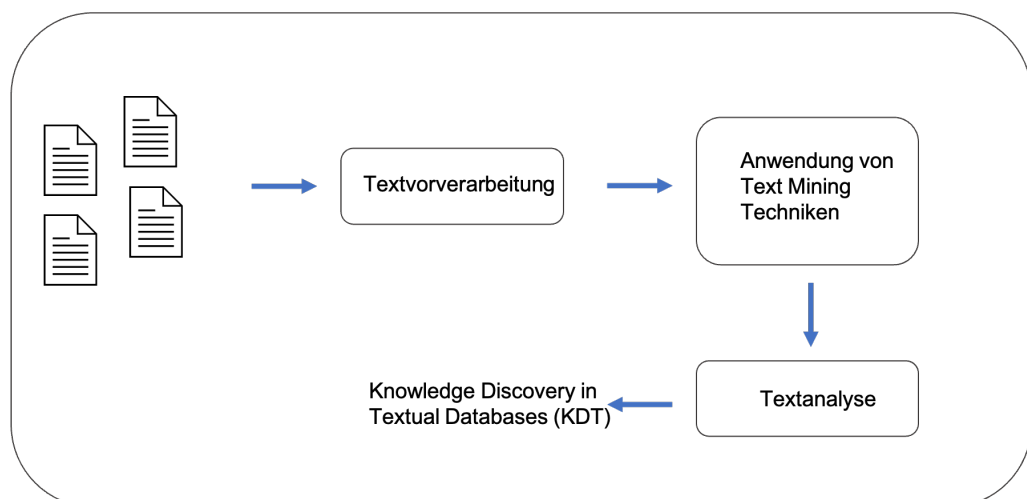
### 2.1 Text Preprocessing

Text Mining bezeichnet den Prozess, um wesentliche bekannte aber auch unbekannte Informationen aus Textdaten zu generieren<sup>6</sup> Die Verarbeitung von unstrukturierten Textdaten wird auch als "Knowledge Discovery in Textual Databases" (KDT) bezeichnet und spielt eine signifikante Rolle in Anwendungsgebieten wie

- Information Retrieval
- Information Extraction
- Natural Language Processing<sup>7</sup>

Im Wesentlichen geht es in allen o. g. Anwendungsgebieten um die Wissen durch das Mining der Texte zu generieren.

**Abbildung 2: Text Mining Prozess**



Quelle: Eigene Darstellung in Anlehnung an Mohan, V., 2015, S. 1

<sup>3</sup> Vgl. Savic, S. et al., 2020, S. 1.

<sup>4</sup> Vgl. ebd., S. 1.

<sup>5</sup> Ebd., S. 1.

<sup>6</sup> Vgl. Mohan, V., 2015, S. 1.

<sup>7</sup> Vgl. ebd., S. 1 f.

Wie der Abbildung 2 entnommen werden kann, stellt die Vorverarbeitung von Volltextdaten bei nahezu jeder Aufgabe im Natural Language Processing (NLP) einen essentiellen und kritischen Schritt dar, da hierbei die fundamentale Basis für die Weiterverarbeitung sowie die Entwicklung der Modelle geschaffen wird.<sup>8</sup> Der Begriff der Textvorverarbeitung umfasst dabei die Anwendung unterschiedlicher Techniken/Methoden, bei denen die Textdokumente für die eigentlichen Zielsetzungen vorbereitet werden. Gängige Techniken für die Vorbereitung der Texte für die nachgelagerten Analysen können folgendermaßen aufgeteilt werden:<sup>9</sup>

- Inhalte Extrahieren und Bereinigen
- Annotationen
- Normalisieren

Zu Beginn der Volltextanalysen stehen häufig die Rohfassungen der Texte zur Verfügung. Diese gilt es im ersten Schritt technisch einzulesen. Hierbei werden auch Daten mit eingelesen, dessen Informationsgehalt gering ist. Beispielhaft zu nennen sind hier HTML tags, Werbung, etc. beim Auslesen einer Website<sup>10</sup> oder Grafiken, ASCII-Codes in PDF-Dateien. Demnach ist das Ziel bei dem **Extrahieren und Bereinigen der Inhalte** die Rohdaten soweit zu säubern, bis sich schließlich die reinen Texte als Resultat ergeben. Nachdem die Texte um die technischen Störfaktoren bereinigt wurden, ist die **Tokenization** eine typische Technik der Textextraktion. In dem Prozess der Tokenisation wird der gesamte zu analysierende Text in einzelne Wörter, Phrasen, Symbole, etc. geteilt. Hierbei wird das Ziel verfolgt die Bedeutung einzelner Wörter innerhalb eines Satzes zu analysieren. Die Tokens dienen nämlich als Eingabewerte für viele weitergehende Prozessschritte.<sup>11</sup> In jedem Text befinden sich Wörter, die wenig Informationsgehalt bei der Textanalyse bieten. Solche Wörter werden auch als **Stop Words** bezeichnet. Beispiele für solche Stop Words sind Artikel oder Präpositionen wie „der“, „die“, „das“, „ein“, „in“, „mit“, etc. Im Analyseprozess stellt jedes unterschiedliche Wort eine eigene Dimension dar. Durch die Entfernung der Stop Words wird somit die Dimensionshöhe reduziert bei gleichzeitiger Beibehaltung des Informationsgehaltes des jeweiligen Satzes/Textes.<sup>12</sup> Neben der klassischen Methode, die Stop Words auf Basis einer vordefinierten Liste zu entfernen, sind diverse mathematische und nicht-mathematische Methoden entwickelt

<sup>8</sup> Vgl. Gurusamy, V., Kannan, S., 2014, S. 2.

<sup>9</sup> Vgl. Pahwa, B., Taruna, S., Kasliwal, N., 2018, S. 1.

<sup>10</sup> Vgl. ebd., S. 1.

<sup>11</sup> Vgl. Gurusamy, V., Kannan, S., 2014, S. 2.

<sup>12</sup> Vgl. Mohan, V., 2015, S. 3.

worden, um Stop Words in Texten zu identifizieren und zu bereinigen.<sup>13</sup>

**Annotationen: POS ergänzen!!!!** Die Annotationen eines Textes sollen die Funktion des jeweiligen Wortes im Kontext des gesamten Satzes identifizieren.

**Normalisieren: Lemma / Stemming** Bei dem **Normalisieren** von Texten wird das Ziel verfolgt ähnliche Wörter zu vereinheitlichen bzw. diese auf einen Standard zu bringen. Dieser Prozess soll vor allem die Dimensionen reduzieren, um die Berechnungen zu vereinfachen und gleichzeitig die Effizienz durch die Standardisierung der Wörter erhöhen.

Bei der Normalisierung von Wörtern wird häufig auf die Techniken des **Stemming** oder der **Lemmatization** zurückgegriffen. Das Stemming ist ein Prozess, der zugrundeliegende Wörter auf den Wortstamm herunterbricht.<sup>14</sup> Dieser Wortstamm ist im Ergebnis häufig kein echtes Wort, sondern oftmals eine Buchstabenkombination bzw. ein Präfix, den viele Wörter gemeinsam haben.<sup>15</sup> Es existiert eine Vielzahl an Stemming-Algorithmen, dessen Performance vom jeweiligen Einsatzbereich abhängt, sodass noch kein Standard etabliert ist.<sup>16</sup>

Die Idee bei der Lemmatization ist das so genannte "Lemma" oder auch die Vokabularform eines Wortes zu identifizieren.<sup>17</sup> Der Prozess ist ähnlich dem des Stemming, jedoch mit dem Unterschied im Ausgabewert. Während beim Stemming der Ausgabewert der Wortstamm ist und oftmals kein echtes Wort, ist der Ausgabewert bei der Lemmatization das Grundwort aus beispielsweise einem Wörterbuch.<sup>18</sup> Die Abbildung 3 verdeutlicht die Unterschiede der Lemmatization und des Stemming anhand des englischen Wortes "change" bzw. dessen Abwandlungen.

---

<sup>13</sup> Detailliertere Informationen zu unterschiedlichen Methoden für die Entfernung von Stop Words können *Mohan, V.*, 2015 entnommen werden.

<sup>14</sup> Vgl. *Khyani, D., B S, S.*, 2021, S. 5.

<sup>15</sup> Vgl. ebd., S. 5.

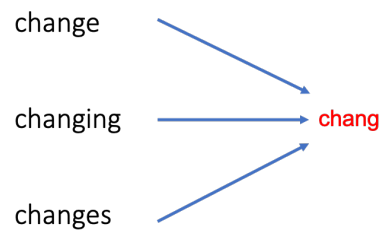
<sup>16</sup> Eine ausführliche Analyse unterschiedlicher Stemming-Algorithmen kann *jivani.2011.* entnommen werden.

<sup>17</sup> Vgl. *Khyani, D., B S, S.*, 2021, S. 7.

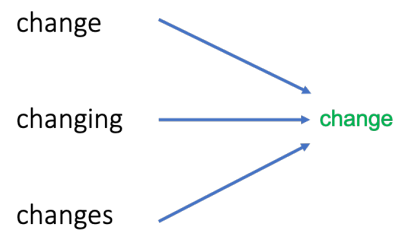
<sup>18</sup> Vgl. ebd., S. 7.

**Abbildung 3: Stemming vs. Lemmatization**

## Stemming



## Lemmatization



*Quelle: Eigene Darstellung in Anlehnung an Khyani, D., B S, S., 2021, S. 7*

## 2.2 Named Entity Recognition

## 2.3 Named Entity Linking

## **3 Praxis**

### **3.1 Markeranalyse**

### **3.2 Überführung in Tabellenstruktur**

### **3.3 Markerkorrelationen**

### **3.4 Aufstellung des Gratingsystems**

## **4 Fazit**

Wünsche Euch allen viel Erfolg für das 7. Semester und bei der Erstellung der Thesis. Über Anregungen und Verbesserung an dieser Vorlage würde ich mich sehr freuen.

## Anhang

### Anhang 1: Beispielanhang

Dieser Abschnitt dient nur dazu zu demonstrieren, wie ein Anhang aufgebaut sein kann.







#### Anhang 1.1: Weitere Gliederungsebene

Auch eine zweite Gliederungsebene ist möglich.

### Anhang 2: Bilder

Auch mit Bildern. Diese tauchen nicht im Abbildungsverzeichnis auf.

#### Abbildung 4: Beispielbild

Name	Änderungsdatum	Typ	Größe
 abbildungen	29.08.2013 01:25	Dateiordner	
 kapitel	29.08.2013 00:55	Dateiordner	
 literatur	31.08.2013 18:17	Dateiordner	
 skripte	01.09.2013 00:10	Dateiordner	
 compile.bat	31.08.2013 20:11	Windows-Batchda...	1 KB
 thesis_main.tex	01.09.2013 00:25	LaTeX Document	5 KB

## Literaturverzeichnis

*Gurusamy, Vairaprakash, Kannan, Subbu* (2014): Preprocessing Techniques for Text Mining, in: o. O., 2014-10-09

*Khyani, Divya, B S, Siddhartha* (2021): An Interpretation of Lemmatization and Stemming in Natural Language Processing, in: Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology, 22 (2021), S. 350–357

*Mohan, Vijayarani* (2015): Preprocessing Techniques for Text Mining - An Overview, in: (2015)

*Pahwa, Bhumika, Taruna, S., Kasliwal, Neeti* (2018): Sentiment Analysis- Strategy for Text Pre-Processing, in: IJCA, 180 (2018), Nr. 34, S. 15–18, [Zugriff: 2021-08-03]

*Savic, S., Leeman, L., El-Shanawany, T., Ellis, R., Gach, J.E., Marinho, S., Wahie, S., Sargur, R., Bewley, A.P., Nakonechna, A., Randall, R., Fragkas, N., Somenzi, O., Marsland, A.* (2020): Chronic Urticaria in the Real-life Clinical Practice Setting in the UK: Results from the Noninterventional Multicentre AWARE Study, in: Clin. Exp. Dermatol. 45 (2020), Nr. 8, S. 1003–1010, [Zugriff: 2021-06-29]

---

## Ehrenwörtliche Erklärung

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist, insbesondere dass ich alle Stellen, die wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen sind, durch Zitate als solche gekennzeichnet habe. Ich versichere auch, dass die von mir eingereichte schriftliche Version mit der digitalen Version übereinstimmt. Weiterhin erkläre ich, dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde/Prüfungsstelle vorgelegen hat. Ich erkläre mich damit **einverstanden/nicht einverstanden**, dass die Arbeit der Öffentlichkeit zugänglich gemacht wird. Ich erkläre mich damit einverstanden, dass die Digitalversion dieser Arbeit zwecks Plagiatsprüfung auf die Server externer Anbieter hochgeladen werden darf. Die Plagiatsprüfung stellt keine Zurverfügungstellung für die Öffentlichkeit dar.

Münster, 4.8.2021

(Ort, Datum)

A handwritten signature in black ink, consisting of a large, stylized 'H' followed by a series of loops and a final flourish.

(Eigenhändige Unterschrift)