Analyse einer Online-Umfrage zur Zufriedenheit mit Finanzämtern mit SAS Contextual Analysis

Nils Hermes
Freie Universität Berlin
Wirtschaftswissenschaften
Garystraße 21
14195 Berlin
nils.hermes@posteo.de

Zusammenfassung

Ziel der vorliegenden Arbeit „Analyse einer Online-Umfrage zur Zufriedenheit mit Finanzämtern mit SAS Contextual Analysis“ war es, Methoden des Natural Language Processing (NLP) auf einen Datensatz anzuwenden. Dazu werden Verfahren des Text Minings, konkret implementiert in dem Software Paket SAS Contextual Analysis, auf die Ergebnisse einer Online-Umfrage zur Zufriedenheit mit Finanzämtern angewandt.
Zuerst wird der Datensatz mit deskriptiven statistischen Methoden charakterisiert. Danach werden die Freitextkommentare der Umfrage mit SAS Contextual Analysis prozessiert und kategorisiert. Diese so erzeugten Kategorien dienen als zusätzliche Kovariaten, um die Zufriedenheit der Befragten im Rahmen eines kumulativen logistischen Modells besser modellieren zu können.
Der Autor konnte zeigen, dass sich Methoden des Text Minings erfolgreich anwenden lassen, um die Auswertung von Texten computergestützt zu beschleunigen und die gewonnenen Erkenntnisse zur verbesserten statistischen Modellierung heranzuziehen.
Die Arbeit wurde von Herrn Prof. Dr. Ulrich Rendtel betreut. Herr Prof. Dr. Frank Hechtner stellte den Zugang zum Datensatz zur Verfügung. SAS Institute GmbH Deutschland stellte im Rahmen eines Stipendiums einen Laptop mit der verwendeten Software zur Verfügung. Damit ist die Arbeit ein Beispiel für die erfolgreiche Kooperation von unterschiedlichen universitären Arbeitsgruppen und von Universität und Wirtschaft im Rahmen eines angewandten Projektes.

Schlüsselwörter: Data Science, Natural Language Processing, Text Mining, SAS Contextual Analysis, Prediction, Cumulative Logistic Model

1 Einleitung

Der vorliegende Text basiert auf der Bachelorarbeit des Autors. Die Bachelorarbeit wurde gekürzt und auf die wichtigsten Teile und Ergebnisse kondensiert. Sie wurde von Herrn Prof. Dr. Ulrich Rendtel und Herrn Prof. Dr. Frank Hechtner betreut. Sie ist in Kooperation mit der SAS Institute GmbH Deutschland [1] im Rahmen eines SAS Student Fellowships entstanden. SAS Institute stellte für die Bearbeitung der Bachelorarbeit einen Laptop bereit. Die Auswertung der Daten erfolgt mit SAS Contextual Analysis, einem webbasierten Text-Mining-Paket, und SAS Studio, einem webbasierten Statistikpaket.

Der Term *Big Data* ist seit einigen Jahren in aller Munde: er beschreibt zusammenfassend die Sammlung, Bearbeitung, Aufbereitung und Interpretation von großen Datenmengen. Akteure auf diesem Feld sind die Wissenschaft, Unternehmen und Regierungen. Ziel in diesem interdisziplinären Gebiet ist es, mit Techniken aus der Informatik, Mathematik und Statistik sog. *Data Mining* zu betreiben, also die Extraktion von Mustern und Zusammenhängen aus den erhobenen Daten. Das sog. *Text Mining* ist eine Unterdisziplin des Data Minings und beschäftigt sich mit der automatisierten Auswertung einer großen Menge von (digitalen) Texten mithilfe von statistischen Verfahren. Das zugehörige Wissenschaftsfeld ist die computergestützte Linguistik (sog. *Computational Linguistics*).

Ziel dieser im Rahmen eines SAS Student Fellowships verfassten Bachelorarbeit ist es, mit der Software SAS Contextual Analysis und der Vorgehensweise des Text Minings einen Datensatz zu analysieren und zu interpretieren. Die Daten werden von Herrn Prof. Dr. Frank Hechtner zu Verfügung gestellt und stammen aus einer Online-Befragung der Finanzverwaltung im Jahr 2016. Hierbei waren Steuerpflichtige aufgerufen, eine Reihe von Parametern der jeweiligen Finanzämter zu bewerten, z.B. die Freundlichkeit der Mitarbeiter. Außerdem wurde ihre Gesamtzufriedenheit mit ihrem jeweiligen Finanzamt abgefragt. Zusätzlich hatten die Befragten die Möglichkeit, in einem Freitextfeld persönliche Rückmeldungen zu verfassen. Die Kommentare sind in der deutschen Sprache verfasst. Diese Kommentare werden mithilfe von Text-Mining-Methoden analysiert und kategorisiert, um in den Kommentaren enthaltene Themen zu entdecken. Die so gewonnenen Kategorien (Themen) werden dann als zusätzliche Kovariate in einem kumulativen logistischen Modell verwendet.

2 Daten

Bei dem Datensatz, der in dieser Arbeit verwendet wird, handelt es sich um eine freiwillige Online-Umfrage zur Zufriedenheit mit der Finanzverwaltung, die im Jahr 2016 unter Steuerpflichtigen durchgeführt wurde. Der Zugang zu den Daten wurde von Herrn Prof. Dr. Frank Hechtner zur Verfügung gestellt.

In diesem Kapitel wird der Datensatz (N = 33.906) vorgestellt (Tab. 1) und die Verteilung der einzelnen Variablen beschrieben. Neben demographischen (kategorialen) Daten wie Alter und Geschlecht sowie Bildung wurden die Zustimmung bzw. Ablehnung zu einzelnen Aussagen gefragt, die den Service der Finanzämter betreffen, z.B. Zufriedenheit (Abb. 1), Freundlichkeit, Hilfsbereitschaft und Verständlichkeit. Diese Daten sind gemäß einer fünfteiligen Likert-Skala kodiert.

Neben der Gesamtzufriedenheit sind die Kommentare der Befragten von größerem Interesse, da diese mithilfe von Text-Mining-Methoden analysiert werden sollen.

Die Kommentare wurden als Antwort auf die Frage „Haben Sie darüber hinaus Anregungen zur Verbesserung der Angebote und Leistungen Ihres Finanzamtes? Oder gibt es Angebote und Leistungen, mit denen Sie besonders zufrieden sind und an denen Ihr Finanzamt / die Finanzverwaltung festhalten sollte?“ verfasst.

Etwa 43% (n = 14.597) der Befragten haben einen Kommentar verfasst (Abb. 1). Die Zeichenlänge der Kommentare ist (natürlich) logarithmisch normalverteilt (Abb. 1).

Der durchschnittliche Kommentar ist 278 Zeichen lang. Damit beläuft sich der Gesamtumfang der Kommentare auf 4.057.966 Zeichen = 14.597 [Kommentare] x 278 [Zeichen pro Kommentar]. Dies entspricht 2.705 A4-Seiten = 4.057.966 [Zeichen] x 1.500^{-1} [Zeichen pro Seite]$^{-1}$ [2].

Tabelle 1: Spezifikation der abgefragten Variablen und ihrer Skalen der Online-Befragung zur Finanzverwaltung aus dem Jahr 2016.

Variable	***Inhalt***	***Skala***
ef1	Ich bin mit meinem Finanzamt generell zufrieden.	Likert
ef6	Die Mitarbeiterinnen und Mitarbeiter begegnen mir freundlich und zuvorkommend.	Likert
ef7	Die Mitarbeiterinnen und Mitarbeiter sind hilfsbereit und unterstützen mich.	Likert
ef10	Die Steuererklärungsvordrucke und die dazugehörigen Erläuterungen sind verständlich.	Likert
ef12	Mein Finanzamt geht kleinlich bei der Prüfung der Steuererklärung(en) vor.	Likert
ef17	Die Bearbeitungszeit meiner Steuererklärung(en) ist angemessen.	Likert
ef23	Ich finde, dass man bei der Steuererklärung ehrlich sein sollte.	Likert
ef24	Ich finde gut, dass die Finanzverwaltung Steuer-CDs ankauft, um damit Steuerhinterziehung zu bekämpfen.	Likert
ef25	Das Informationsangebot der Finanzverwaltung bietet mir ausreichend Möglichkeiten, mich über steuerliche Änderungen zu informieren.	Likert
ef30	Wie haben Sie Ihre letzte Steuererklärung an das Finanzamt übermittelt?	Kategorien
ef32	Alter	Kategorien
ef33	Geschlecht	Kategorien
ef38	Wie würden Sie Ihr steuerliches Wissen einschätzen?	Kategorien
ef39	Welches ist Ihr höchster erreichter Schulabschluss?	Kategorien
zeit1	Gesamtzeit für die Bearbeitung der Umfrage (in Sekunden)	Metrisch
kommentar	Haben Sie darüber hinaus Anregungen zur Verbesserung der Angebote und Leistungen Ihres Finanzamtes? Oder gibt es Angebote und Leistungen, mit denen Sie besonders zufrieden sind und an denen Ihr Finanzamt / die Finanzverwaltung festhalten sollte?	Freitext

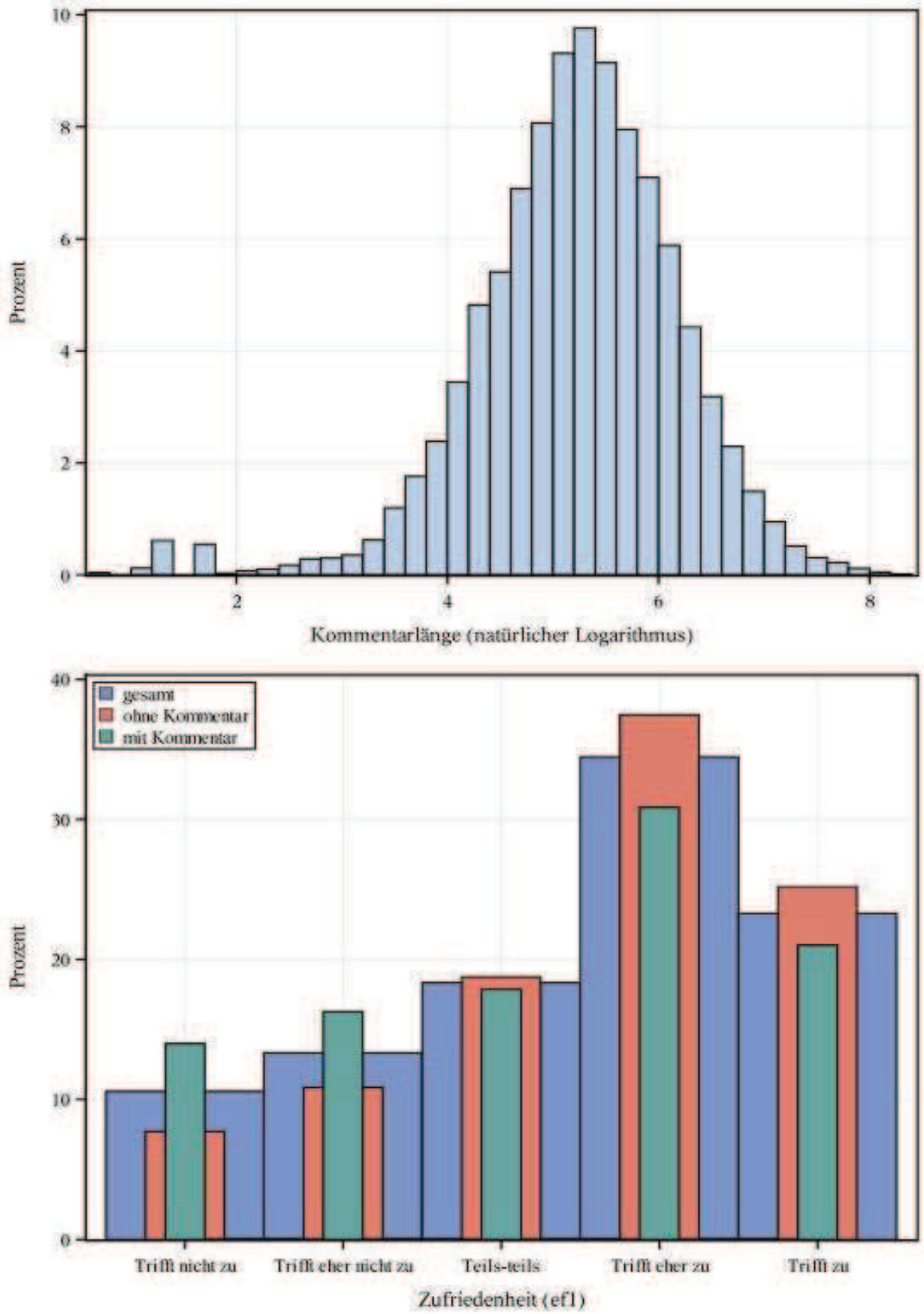

Abbildung 1: Übersicht über die generelle Zufriedenheit der Befragten, gruppiert nach Existenz eines Kommentars und Verteilung der Zeichenlänge der Kommentare

3 Methoden

In diesem Kapitel werden die verwendeten Methoden diskutiert. In diesem Kapitel wird die Text-Mining-Methode erläutert, sie ist in Form der Prozedur HPTMINE (High-Performance Text Mining) in der SAS Software implementiert und wird auch von SAS Contextual Analysis verwendet.

3.1 Text Mining

Der folgenden Erläuterungen der Text-Mining-Methode und anschließenden Themen-Extraktion sind hauptsächlich aus folgenden Quellen entnommen: dem Paper „A practical guide to text mining with topic extraction“ von Karl et al. [3] und Kapitel 13 aus dem Buch „Data Mining - The Textbook“ von Aggarwal [4]. Die allgemeine Vorgehensweise ist hier schematisch aufgelistet:

- Knappe Beschreibung der zu lösenden Problemstellung.
- Sammlung von passenden Text- und strukturierten Daten.
- Verarbeitung und Filtern der Texte durch die Entfernung von Wörtern und Zeichen wie Rechtschreibfehler, allgemeinen Wörtern (Stoppwörter), irrelevanten Wörtern, die aus vorheriger Verarbeitung stammen, seltene Wörter, die nur in wenigen Dokumenten auftauchen, Wörter, die zu kurz oder zu lang sind, Zahlen und Nicht-Standardzeichen. Zusätzlich kann es manchmal sinnvoll sein, Synonyme unter einem einzelnen, repräsentativen Wort zusammenzufassen.
- Transformation der Texte in eine gewichtete Matrix, um weitere statistische Analysen durchführen zu können.
- Erkundung und Entdeckung von Themen und Gemeinsamkeiten.
- Gruppierung von ähnlichen Dokumenten und Wörtern.
- Erzeugung neuer strukturierter Variablen aus Text, die für weitere Vorhersageanalyse (sog. *predictive analysis*) verwendet werden können.

Textverarbeitung

Bevor Themen und andere strukturierte Variablen erzeugt werden können, muss die Menge von Texten (sog. *Dokumenten*, in ihrer Gesamtheit sog. *Korpus*), die ihrerseits Wörter (sog. *Terme*) enthalten, eingelesen werden. Terme sind die zugrundeliegenden Objekte, deren Oberflächenerscheinung sich als Wort oder Wortgruppe in einer Sprache äußert. Zusätzlich muss der Korpus in eine zweckmäßige Form gebracht und bereinigt werden.

Dokument-Term-Matrix

In einem ersten Schritt wird die Menge der Dokumente in eine Matrix überführt: die Dokument-Term Matrix (DTM). Diese Matrix lässt sich in der folgenden Art und Weise darstellen:

$$a_{i,j} = f_{i,j} = \text{Anzahl bzw. Frequenz des Terms } i \text{ im Dokument } j$$

Im Allgemeinen ist diese Matrix **A** eine dünn besetzte Matrix (sog. *sparse matrix*): viele ihrer Elemente $a_{i,j}$ sind null. Die aus den Kommentaren der Umfragen konstruierte Matrix enthält 479.394.674 Elemente = 14.597 [Dokumente] × 32.842 [Terme].

Nun ist es zweckmäßig, die Dimensionen dieser Matrix zu reduzieren und Terme durch linguistische Sprachverarbeitung und andere Methoden zusammenzuführen: Ziel ist es, sprachlich gleiche Terme zusammenzufassen sowie Terme zu entfernen, die wenig bis gar keine Information enthalten. Mehrere Terme können unter einem Elternterm zusammengefasst werden, indem sie auf ihre sprachliche Normalform reduziert werden (sog. *stemming*), ihre Frequenzverteilung im Korpus verglichen wird (sog. f*requency matching*) und/oder sie durch manuelle Wortlisten subsumiert werden (sog. *Synonymlisten* und *Wortgruppenlisten*). Die Zahl der Terme kann weiter reduziert werden, indem Terme mit hohen Frequenzen entfernt werden (z.b. Präpositionen, Personalpronomen, Füllwörter), Terme mit niedrigen Frequenzen entfernt werden (sog. *cutoff frequency*) und/oder sie durch manuelle Wortlisten (sog. *Stoppwörterlisten*) entfernt werden.

Als Basis für die verwendete Stoppwörterliste diente eine Liste, die etwa 600 Wörter enthielt [5]. Zusätzlich wurden zu dieser Liste manuell weniger als 100 Wörter hinzugefügt. Die Standardliste des SAS Text Miners SASHELP.GRMNSTOP ist mit ca. 100 Einträgen etwas knapp bemessen.

Diese nun reduzierte DTM wird dann noch weiter in ihren Dimensionen durch eine Singulärwertzerlegung (sog. *singular value decomposition*) reduziert.

Reduzierte DTM

Nachdem die obigen Schritte ausgeführt sind, erhält man die reduzierte DTM. Von den 32.842 Termen verbleiben 5.359 Terme, eine Reduktion um > 83%. Damit enthält die resultierende Matrix 78.225.323 Elemente = 14.597 [Dokumente] × 5.359 [Terme].

3.2 Konstruktion von Themenbereichen

Die reduzierte DTM ist der Startpunkt für die nächste Abfolge von Verarbeitungsschritten: Ziel ist es, Themenbereiche aus dem Korpus zu konstruieren. Themenbereiche sind Mengen von Termen, denen jeweils eine Menge von Dokumenten zugeordnet wird. Hierzu wird die reduzierte DTM mittels einer Singulärwertzerlegung (SVD) bearbeitet. Die SVD wird verwendet, um die Dimensionalität der DTM weiter zu verringern und Themen zu konstruieren.

Termgewichtung

Bevor die Singulärwertzerlegung ausgeführt wird, können die einzelnen Termfrequenzen $f_{i,j}$ der DTM A gewichtet werden. Es ist im Allgemeinen sinnvoll, Terme, die weniger häufig auftreten, stärker zu gewichten. Die Gewichtung ist hauptsächlich empirischer Natur. SAS Contextual Analysis benutzt die Prozedur HPTIME: die hier aufgeführten Termgewichte stammen aus der Dokumentation dieser Prozedur.

Die gewichtete Termfrequenz $f_{i,j}$ setzt sich aus einer Termgewichtungsfunktion *w* und einer Zellgewichtungsfunktion *g* zusammen:

$$f_{i,j,gew} = w(f_{i,j}) \times g(f_{i,j})$$

Die Zellgewichtungsfunktion $w_{i,j}$ ist eine Logarithmusfunktion, die hohe Argumente „bestraft“:

$$g_{i,j} = \log_2(f_{i,j} + 1)$$

Für die Termgewichte $w_{i,j}$ wird die folgende Funktion verwendet. Ihre Struktur leitet sich aus der sog. *informationstheoretischen Entropie* ab. $p_{i,j}$ ist die Wahrscheinlichkeit, dass der Term *i* im Dokument *j* auftaucht und kann aus seinen Frequenzen $f_{i,j}$ approximiert werden. *n* ist die Menge aller Dokumente.

$$w(f_{i,j}) = 1 + \sum_i \frac{p_{i,j} \log_2(p_{i,j})}{\log_2 \mathrm{n}} \text{ mit } p_{i,j} = \frac{f_{i,j}}{\sum_i f_{i,j}}$$

Die gewichtete reduzierte Dokument-Term Matrix wird nun einer Singulärwertzerlegung unterzogen.

Singulärwertzerlegung

Die folgende Beschreibung der Singulärwertzerlegung ist von Alter et al. entnommen [6]. Als Singulärwertzerlegung bezeichnet man das Produkt einer $m \times n$ Matrix M der folgenden Form (A^T bezeichnet den Vorgang von Transponierung einer Matrix A):

$$M = U\Sigma V^T$$

Da die Matrix $U(V)$ orthogonal ist, ist das Produkt $U^T U(V^T V)$ die Identitätsmatrix I. Die Spalten der Matrix $U(V)$ setzen sich aus den Eigenvektoren der Matrix $MM^T(M^T M)$ zusammen. Die Matrix Σ setzt sich aus der Quadratwurzel der Eigenwerte von $MM^T(M^T M)$ zusammen. Ein Algorithmus für die numerische Bestimmung der Singulärwerte befindet sich in [7]. Für eine Dimensionsreduktion der reduzierten DTM kann nun ein Wert k < n verwendet werden, sodass nur die ersten k Singulärwerte verwendet werden. Damit kann die Menge der Themen gesteuert werden, die durch die spätere Projektion erzeugt wird. Die Singulärwerte der DTM sind in Abb. 2 für k = 200 aufgetragen.

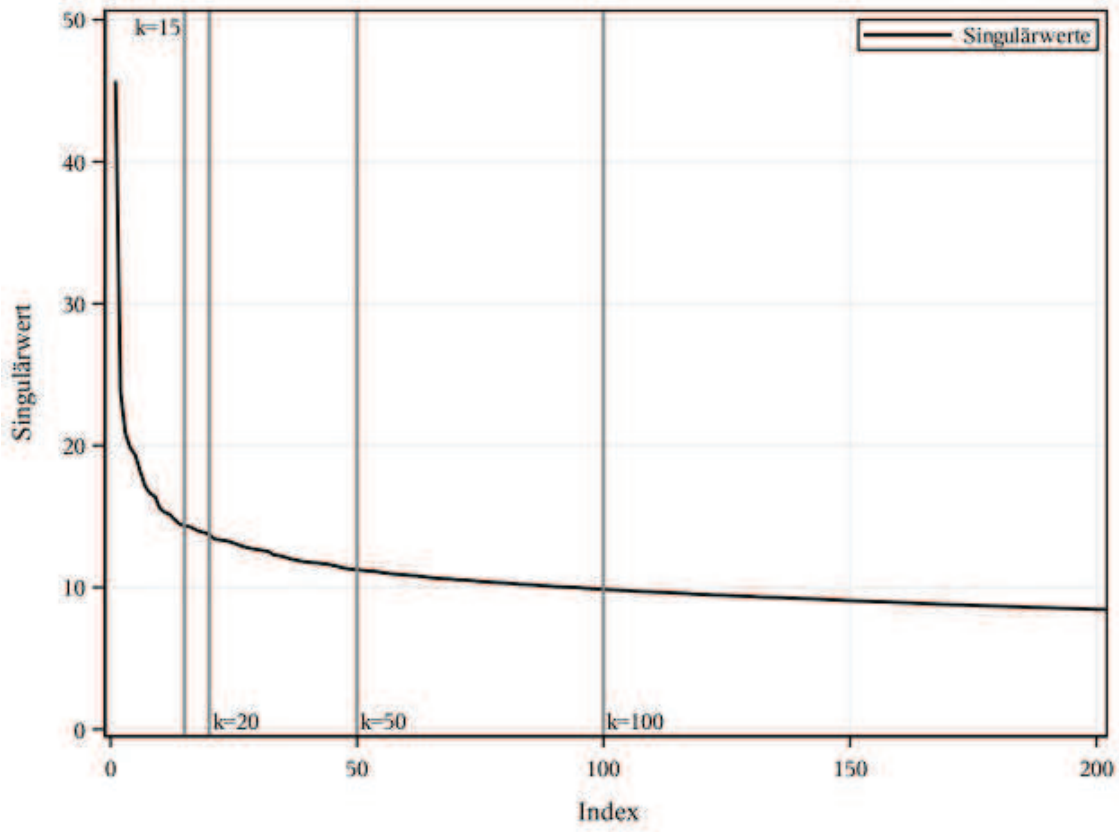

Abbildung 2: Plot der Singulärwerte *k* über ihren Index *i* der Dokument-Term Matrix

Die Auswahl eines Wertes für k ist ein Problem und muss an das entsprechende Vorhaben angepasst werden: es muss zwischen einer größeren Anzahl von Themen und besserer Interpretierbarkeit abgewogen werden. Als Standard wurde k = 15 gewählt. Die Menge der Themen k wird dynamisch im Rahmen der Prozedur HPTMINE bestimmt. Allerdings wurden auch die Ergebnisse für größere, manuell eingestellte Werte betrachtet (k = 20, 50, 100).

Projektion der SVD

Themen sind durch eine Menge von gewichteten Termen charakterisiert. Dokumente mit vielen dieser Terme sind stärker mit einem Thema verbunden als Dokumente mit wenigen dieser Terme. U charakterisiert die Verbindung von Themen und Dokumenten und V charakterisiert die Verbindung von Themen und Termen. Die Themen werden als Rotation der SVD Dimensionen berechnet, um die Quadratsumme der Termbeladung der Matrix V zu maximieren. Hierbei besitzt jedes Thema einen höheren Erklärungsgehalt. Eine Beschreibung dieser Projektion findet sich in der Dokumentation der PROC HPTMINE.

4 Ergebnisse

4.1 Themen

Um die maschinell erzeugten Themen verstehen und analysieren zu können, wurden alle Themen in einen Satz in „menschlicher“ Sprache „zurückübersetzt“. Als Grundlage für diese Sätze diente eine Kombination aus den Termen mit den höchsten Gewichten und die dem Thema zugeordneten Dokumenten mit den höchsten Gewichten.

Einige Themen sind exemplarisch in Form von Wortwolken in Abb. 3 und 4 visualisiert. Hierbei entspricht die Größe der Terme dem Gewicht des entsprechenden Terms im Thema.

Abbildung 3: Wortwolke des Themas 3: „Lange / Schnellere Bearbeitungszeiten. Bearbeitung dauert Monate.“

Abbildung 4: Wortwolke des Themas 4: „Telefonische Erreichbarkeit (der Telefonzentrale) verbessern.“

Die folgenden Themensätze sind für k = 15 Themen konstruiert (Tab. 2). Zusätzlich werden die Themensätze den manuell erzeugten Kategorien zugeordnet. Diese Themen werden in der statistischen Modellierung als zusätzliche Kovariaten zur Erklärung der Gesamtzufriedenheit der Befragten verwendet. Jedem Dokument ist hierbei ein Gewicht zugeordnet, das die Zugehörigkeit zu dem jeweiligen Thema charakterisiert.

Tabelle 2: Übersicht aller Themen mit k = 15 durch Themensätze und Zuordnung zu manuell erzeugten Kategorien.

ID	***Themensatz***	***Kategorie***
1	Verständliche(re) Erläuterungen zum Steuerbescheid.	Verständlichkeit Schriftverkehr / Vordrucke
2	Persönlicher Kontakt zum zuständigen Sachbearbeiter.	Kommunikationsweg
3	Lange/Schnellere Bearbeitungszeiten. Bearbeitung dauert Monate.	Bearbeitungszeit
4	Telefonische Erreichbarkeit (der Telefonzentrale) verbessern.	Telefonische Erreichbarkeit Organisation des FAs
5	ElsterOnline / -Formular ist gut. (Programm auch für andere Betriebssysteme).	Elster
6	Steuerzahler (allgemein). Anschreiben Herr Dr. Walter-Borjans.	Anschreiben
7	Finanzamt / -verwaltung / -beamte (allgemein).	
8	Verständlichere / einfachere Steuererklärungsformulare. kein(e) Beamtendeutsch / -sprache.	Verständlichkeit Schriftverkehr / Vordrucke
9	Freundliche / hilfsbereite / kompetente Mitarbeiter. Zufrieden mit dem Finanzamt.	Freundlichkeit Kompetenz
10	Längere Öffnungs- / Sprechzeiten für Arbeitnehmer / Berufstätige. Mind. einen Tag in der Woche länger offen.	Öffnungszeiten
11	Vereinfachung der/des Steuererklärung / -systems / -rechts. Steuererklärung auf einem Bierdeckel. (kein Steuerberater nötig).	Steuersystem
12	Elektronische (Online-)Einreichung / Übermittlung der Steuererklärung / -belege.	Abgabepflicht
13	(Bearbeitung der) Steuererklärung dauert Monate.	Bearbeitungszeit
14	Persönliche Abgabe (der Steuererklärung) im Bürgerbüro. Bürgerbüro ist gut.	Abgabepflicht Bürgerservice
15	Persönlicher Ansprechpartner für Rückfragen.	Kommunikationsweg

Die folgenden Themensätze sind aus einem Durchlauf mit k = 100 erzeugt worden (s. Tab. 3). Hier wird allerdings nur eine Auswahl von Themen dargestellt, da bei größeren k die Verständlichkeit der Themen im Allgemeinen geringer wird bzw. die Themen nur

noch einzelne Terme enthalten. Daher sollten diese Themen nur explorativ verstanden werden und werden nicht als zusätzliche Kovariaten verwendet.

Tabelle 3: Übersicht einer Auswahl von Themen mit k = 100 durch Themensätze und Zuordnung zu manuell erzeugten Kategorien.

Themensatz	***Kategorie***
Erläuterungen in größerer Schrift.	Verständlichkeit Schriftverkehr / Vordrucke
Belegeinreichung als .pdf-Datei.	Abgabepflicht
Elster-Software für andere Betriebssysteme (MacOS, Linux).	Elster
Emailadresse für Kommunikation.	Kommunikationsweg
Übernahme der Daten aus dem Vorjahr.	Elster
Verlängerung der Abgabefristen.	Prüfung
Steuersystem / Steuerrecht ist (zu) kompliziert.	Steuersystem
Pauschale Anerkennung von Werbe- / Fahrtkosten.	Prüfung
Statusübermittlung online statt per Post.	Rückmeldung
(Unmut über) Teilzeitkräfte.	Organisation des FAs
Eingangsbestätigung der Unterlagen.	Rückmeldung
Weiter so!	
Steuergerechtigkeit: Normale Bürger vs. Großunternehmen / Topverdiener.	Steuergerechtigkeit
Größere Briefkästen.	örtliche Erreichbarkeit
Mehr Personal.	Organisation des FAs
Formulare in Papierform für ältere Menschen.	Abgabepflicht
Rückerstattung soll zügiger erfolgen.	
Gleichstellung von Mann und Frau in den Formularen.	

4.2 Statistische Modellierung

Ziel ist es, den Grad der Zustimmung zu der Variable *efl* „Ich bin mit meinem Finanzamt generell zufrieden.“ durch die anderen Variablen *ef* und die extrahierten Themen *topic_raw* aus den Kommentaren zu modellieren. Der Inhalt der einzelnen Variablen *ef* ist in Kap. 2 beschrieben. Die hier vorgestellten Modelle werden mit der Methode PROC LOGISTIC erstellt. Sie implementiert das kumulative logistische Modell.

Das `class` Statement kodiert die ordinalen bzw. kategorialen Variablen in Effektkodierung, hierbei dient der Mittelwert der jeweiligen Variablen als Referenz. Die Themen liegen als metrische Variablen vor. Basierend auf der obigen Prozedur werden vier Modelle gerechnet, deren Spezifikation findet sich in Tab. 4. Aus Platzgründen wird nur ein Modell aufgelistet. Die Bewertung der Modelle erfolgte auf Basis ihres AIC.

```
1 proc logistic data = survey descending;
2   class
3     ef6 ef7 ef10 ef12 ef17
4     ef23 ef24 ef25 ef30 ef32
5     ef33 ef38 ef39 / desc;
6   model
7     ef1 =
8       ef6 ef7 ef10 ef12 ef17
9       ef23 ef24 ef25 ef30 ef32
10      ef33 ef38 ef39 zeit1 log_lc_n
11      (topic_raw1-topic_raw15);
12   output
13     out = results
14     predicted = prob
15     (selection = stepwise);
16 run;
```

Tabelle 4: Verwendete Variablen in den vier Modellen. Die Optimierung der Modelle erfolgt durch die schrittweise Auswahl (`selection = stepwise`).

Name	***Variablengruppe***
Grundmodell	*ef*
volles Modell	*ef* + *topic_raw*
optimiertes Grundmodell	Untermenge von *ef*
optimiertes volles Modell	Untermenge von *ef* + *topic_raw*

Bevor die Modelle im Einzelnen vorgestellt werden können, muss eine heuristische Beschreibung der Punktschätzer gefunden werden, da durch die große Menge an Variablen und ihrer Freiheitsgrade (*DF*) viele Punktschätzer zustande kommen. Das Grundmodell enthält zum Beispiel im Ergebnis 54 Koeffizienten und damit auch Punktschätzer. Es soll in dieser Arbeit nicht um den exakten Wert der Koeffizienten bzw. Punktschätzer gehen, sondern um die Richtung der Punktschätzer der jeweiligen Variablen und um den zusätzlichen Erklärungsgehalt der Themen. Trotzdem wird die Größe der Punktschätzer grob klassifiziert. Die Variablen mit einer Likert-Skala und die kategorialen Variablen haben *DF + 1* Ausprägungen. Die Anzahl der für jede Variable berechneten Koeffizienten ist *DF*, da sie sich auf eine Referenzausprägung beziehen: der Koeffizient β und sein Punktschätzer $exp(\beta)$ von z.B. *ef17* (Bearbeitungszeit) gibt hierbei die Änderung des Chancenverhältnisses zugunsten größerer Zufriedenheit ausgehend von der niedrigsten Ausprägung von *ef17* (β_0) an.

Die Variablen der Themen *topic_raw* sind metrisch skaliert und haben demnach nur eine Ausprägung. Ihr Wertebereich liegt im Intervall [0,1].

Die Werte der Punkteschätzer für jede Variable sind mit einem Symbol klassifiziert. Die Beschreibung der Symbole ist in Tab. 5 aufgeführt. Die Symbole für die Punktschätzer sind so zu verstehen, dass die Mehrheit der Punktschätzer der jeweiligen Variablen

oberhalb des in der Tabelle aufgeführten Schwellenwertes liegen. Die Punktschätzer innerhalb der nichtmetrischen Variablen sind in den meisten Fällen monoton. Die Trennung der Punktschätzer nach Umfrage und Themen folgt aus den unterschiedlichen Intervallen, auf denen sich ihre Variablen bewegen.

Die Kovariaten in den Tabellen zu jedem Modell sind nach ihrem Erklärungsgehalt (Waldsches χ^2) sortiert.

Tabelle 5: Symbole und Bedeutungen der Punktschätzer der einzelnen Variablen. Die Symbole beschreiben den Einfluss der jeweiligen Variablen auf die Abhängige *ef1* (Zufriedenheit).

Symbol	***Einfluss***	***Punktschätzer (ef)***	***Punktschätzer (topic_raw)***
+ + +	Stark positiv	> 2,00	> 10,0
+ +	Positiv	> 1,50	> 5,0
+	Schwach positiv	> 1,00	> 2,0
(+)	Sehr schwach negativ		> 1,0
0	Kein Einfluss	~ 0,00	~ 1,0
(–)	Sehr schwach negativ		< 1,0
–	Schwach negativ	< 1,00	< 0,5
– –	Negativ	< 0,67	< 0,2
– – –	Stark negativ	< 0,50	< 0,1

4.3 Das optimierte volle Modell

Das optimierte volle Modell enthält alle Variablen *ef*, die Bearbeitungszeit der Umfrage *zeit1* sowie die logarithmierte Kommentarlänge *log_lc_n* sowie alle Themen *topic_raw*. Das Modell verwendet n = 17.056 Beobachtungen aus einer Grundgesamtheit von N = 33.906 Beobachtungen. Alle Kovariaten sind signifikant ($p < 0,05$).

Tabelle 6: Spezifikation aller Variablen und ihrer Punktschätzer des optimierten vollen Modells.

ID	***Name***	***DF***	***Schätzer***
ef17	Bearbeitungszeit	4	+ + +
ef7	Hilfsbereitschaft	4	+ + +
ef6	Freundlichkeit	4	+ + +
ef12	Kleinliche Prüfung der SE	4	– –
ef10	Verständlichkeit	4	+ +
ef25	Informationsangebot	4	+ +
ef24	Ankauf von Steuer-CDs	4	+
log_lc_n	log. Kommentarlänge	1	– –
topic_raw_4	tel. Erreichbarkeit	1	– – –
topic_raw_5	Elster	1	+ + +

ef32	Alter	6	–
topic_raw_13	Bearbeitungszeit	1	– – –
topic_raw_9	freundliche Mitarbeiter	1	+ +
ef30	Art der Übermittlung	3	+
topic_raw_14	persönliche Abgabe	1	+
ef23	Ehrlichkeit	4	+ +
topic_raw_11	Vereinfachung der SE	1	+
topic_raw_8	Verständlichere Formulare	1	+

Vergleicht man das optimierte Grundmodell und das optimierte volle Modell aufgrund des Likelihood-Quotienten-Tests und des AIC, so wird deutlich: das Hinzufügen von Themen liefert einen zusätzlichen Erklärungsgehalt für die Beschreibung der Gesamtzufriedenheit.

4.4 Diskussion

Die Variablen *ef6* „Freundlichkeit der Mitarbeiter“, *ef7* „Hilfsbereitschaft der Mitarbeiter“, *ef10* „Verständlichkeit der Steuererklärung“, *ef12* „Kleinliche Prüfung der SE“, *ef17* „Bearbeitungszeit der SE“ und *ef25* „Informationsangebot“ haben einen hohen Erklärungsgehalt und einen starken positiven Zusammenhang mit der Gesamtzufriedenheit.

Die Variablen *ef23* „Ehrlichkeit bei der SE“ und *ef24* „Ankauf von Steuer-CDs“ haben einen geringeren Erklärungsgehalt, sind aber trotzdem positiv mit der Gesamtzufriedenheit verknüpft.

Die sozio-demographischen Faktoren *ef32* „Alter“, *ef33* „Geschlecht“ und *ef39* „Schulabschluss“ sowie *ef38* „Steuerliches Wissen“ haben einen sehr geringen bis nicht signifikanten Erklärungsgehalt der Gesamtzufriedenheit.

Die Bearbeitungszeit der Umfrage *zeit1* spielt keine Rolle für die Gesamtzufriedenheit. Die logarithmierte Länge der Kommentare *log_lc_n* spielt eine negative Rolle für die Gesamtzufriedenheit: Wer einen Kommentar verfasst, ist unzufriedener mit der Arbeit des Finanzamtes. Die Themen *topic_raw4* „Telefonische Erreichbarkeit (der Telefonzentrale) verbessern.“ und *topic_raw13* „(Die Bearbeitung der) Steuererklärung dauert Monate.“ sind stark negativ konnotiert.

Die Themen *topic_raw5* „ElsterOnline / -programm ist gut.“ und *topic_raw9* „Die Mitarbeiter sind freundlich/hilfsbereit/kompetent.“ sind stark positiv konnotiert.

Die Themen *topic_raw8* „Verständlichere / Einfachere Steuererklärungsformulare / kein(e) Beamtendeutsch/-sprache.“, *topic_raw11* „Vereinfachung der Steuererklärung (kein Steuerberater nötig) / (SE auf einem) Bierdeckel.“ und *topic_raw14* „Persönliche Abgabe (der SE / Belege) im Bürgerbüro / Bürgerbüro ist gut.“ sind positiv konnotiert.

5 Zusammenfassung

Im Rahmen dieser Arbeit wurden die Methoden des Text Minings erfolgreich für die Kategorisierung von Freitextkommentaren einer Online-Umfrage zur Zufriedenheit mit Finanzämtern angewendet. Diese algorithmische Methode ist im Vergleich zu einer manuellen Kategorisierung von großen Textmengen sehr schnell und damit durchaus auch ökonomisch sinnvoll. Allerdings ist es erforderlich, dass die nötigen Routinen für das Text Mining bereits als Softwarepaket vorliegen bzw. manuell implementiert werden können. Als Vorbereitung für die Singulärwertzerlegung muss eine passende Stoppwörterliste für das entsprechende Themengebiet gefunden werden. Hierbei kann auf bereits vorhandene Listen zurückgegriffen werden oder eine Liste manuell angelegt werden. Auch ist eine manuelle Überprüfung der resultierenden Termlisten nötig: müssen zusätzliche Terme entfernt bzw. aufgenommen werden?

Es wurde gezeigt, dass die maschinell extrahierten Themen in guter Übereinstimmung mit den manuell erzeugten Kategorien sind. Durch den Parameter k, die Menge der konstruierten Themen, kann die Spezifität der Themen gesteuert werden. Es wurde demonstriert, dass sowohl häufig auftretende und damit wichtige Anliegen als auch spezifische, kleinere Anliegen erfasst werden können. Diese Themen können dann den Finanzämtern, in eine Handlungsanweisung übersetzt, zur Verbesserung ihres Dienstleistungsangebots übergeben werden. An dieser Stelle wäre auch denkbar, ein automatisches Beschwerdemanagementsystem zu implementieren, das schriftliche Beschwerden oder Hinweise aufgrund bereits erzeugter Kategorien den entsprechenden Stellen im Finanzamt zuweist. Es wurde demonstriert, dass die erzeugten Themen als zusätzliche Kovariaten zur besseren Beschreibung der Gesamtzufriedenheit mit den Finanzämtern herangezogen werden können. Als zweckmäßiges statistisches Modell wurde hierfür das kumulative logistische Modell verwendet. Eine erwähnenswerte Eigenschaft der Kommentare ist, dass die Länge der Kommentare logarithmisch normalverteilt ist. Bereits die Tatsache, dass, wenn ein Kommentar verfasst wurde, die Zufriedenheit des Verfassers geringer ist, ist hilfreich.

Aufgrund zeitlicher Beschränkungen konnte die detaillierte Analyse der Stimmungen der Befragten im Rahmen einer Stimmungsanalyse (sog. *sentiment analysis*) nur rudimentär durchgeführt werden: die Ergebnisse sind nicht in dieser Arbeit aufgeführt. Dies wäre ein Feld für zukünftige Forschungsvorhaben.

Abschließend ist zu sagen, dass die Methoden des Text Minings erfolgreich in die Analyse von Umfragen eingebunden werden konnten und damit einen organischeren Blick auf die Anliegen der Befragten geben.

Literatur

[1] SAS Institute (2017). *SAS Institute GmbH*. URL:https://www.sas.com/de_de/contact.html (besucht am 23. 10. 2017).

[2] VG Wort (2017). *Zeichenzahl pro Seite*. URL:http://www.vgwort.de/-verguetungen/auszahlungen/wissenschaftliche-publikationen/fach-und-sachbuecher.html (besucht am 18. 09. 2017).

[3] A. Karl, J. Wisnowski und W. H. Rushing (2015). „A practical guide to text mining with topic extraction“. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 7(5), S. 326–340. URL: http://dx.doi.org/10.1002/wics.1361

[4] C. C. Aggarwal (2015). *Data mining: the textbook*. Springer.

[5] M. Götze, S. Geyer (2017). *Stoppwörterliste*. URL: https://solariz.de/de/downloads/6/german-enhanced-stopwords.htm (besucht am 29. 08. 2017).

[6] O. Alter, P. O. Brown und D. Botstein (2000). „Singular value decomposition for genome-wide expression data processing and modeling“. In: *Proceedings of the National Academy of Sciences* 97(18), S. 10101–10106. EPrint: http://www.pnas.org/content/97/18/10101.full.pdf URL: http://www.pnas.org/content/97/18/10101.abstract

[7] G. Golub, W. Kahan (1965). „Calculating the singular values and pseudoinverse of a matrix“. In: *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 2(2), S. 205–224.