Grundlagen: Methodenbericht | August 2021

# Befristungen in der Statistik der gemeldeten Arbeitsstellen

Analysen mit Text Mining

Bundesagentur für Arbeit
Statistik

# Impressum

**Produktlinie/Reihe:** Grundlagen: Methodenbericht

**Titel:** Befristungen in der Statistik der gemeldeten Arbeitsstellen - Analysen mit Text Mining

**Veröffentlichung:** August 2021

**Herausgeberin:** Bundesagentur für Arbeit
Statistik/Arbeitsmarktberichterstattung

**Autoren:** Arsen Çelikel
Matthias Gehricke
Marlene Hellmann
Dr. Jörg Szameitat

**Rückfragen an:** Konzepte und Methoden der Statistik, Fachliche Entwicklung
Regensburger Straße 104
90478 Nürnberg

**E-Mail:** Zentraler-Statistik-Service@arbeitsagentur.de

**Telefon**: 0911 179-3632

**Fax:** 0911 179-1131

**Weiterführende statistische Informationen:**

**Internet:** http://statistik.arbeitsagentur.de

**Zitierhinweis:** Statistik der Bundesagentur für Arbeit,
Grundlagen: Methodenbericht – Text Mining – Ein explorativer Ansatz der Methode des Text Mining auf die gemeldeten Stellen im Bereich Befristung, Nürnberg, August 2021

# Inhaltsverzeichnis



**Abkürzungsverzeichnis**

| | |
|---|---|
| BA | Bundesagentur für Arbeit |
| DWH | DataWarehouse |
| IAB | Institut für Arbeitsmarkt- und Berufsforschung |
| IE | Informationsextraktion |
| NLP | Natural Language Processing |
| VerBIS | Vermittlungs- Beratungs- und Informationssystem |

**Abbildungsverzeichnis**



**Tabellenverzeichnis**

# 0 Kurzfassung

Die Statistik der Bundesagentur für Arbeit (BA) erprobt erstmals Methoden des Text Mining für die Analyse von Stellenanzeigentexten mit dem Ziel, die vorhandenen statistischen Daten zu prüfen und ggf. anzureichern. Der vorliegende Methodenbericht stellt die angewandte Methodik vor, mit der systematisch die Validität des Merkmals „Befristung“ untersucht wurde, zeigt die erzielten Ergebnisse auf und gibt einen Ausblick, wie die gewonnenen Erkenntnisse in die statistische Verarbeitung miteinbezogen werden können.

Das hier beschriebene Vorhaben erprobt Text Mining erstmals anhand einer breiten Basis amtlicher Daten. Verwendet wird dieselbe Datenquelle, aus der auch die Statistik der gemeldeten Arbeitsstellen erstellt wird, angereichert um die Ausschreibungstexte der Stellenangebote. Diese werden bislang nicht für die statistische Berichterstattung der BA genutzt. Amtliche Statistikdaten als Grundlage für Text Mining heranzuziehen hat mehrere Vorteile gegenüber Web Scraping mit Online-Stellenbörsen. So hat die Statistik der BA bei Text Mining mit gemeldeten Arbeitsstellen zum einen profunde Kenntnis über Struktur und Historizität der Datenquelle und zum anderen sind alle Merkmale verfügbar, die auch in den amtlichen Statistikdaten enthalten sind – und nicht nur Texte.

Mit den um Stellenanzeigentexte erweiterten Statistikdaten wurden unterschiedliche Modelle und Algorithmen getestet. Aufgrund der Ergebnisse, ihrer guten Interpretierbarkeit und der hohen Modellgüte fiel die fachliche Entscheidung auf den sog. XGBoost-Klassifikator. Das Modell könnte die statistische Information, ob ein Stellenangebot befristet oder unbefristet ausgeschrieben ist, gegenüber dem jetzigen Stand verbessern.

Das Befristungsmodell wird derzeit einem einjährigen, manuellen Testbetrieb unterzogen, um die Integration in die regelmäßige Statistikproduktion vorzubereiten. Sobald die Ergebnisse hierzu vorliegen, ist – neben der Entscheidung über eine statistische Veröffentlichung als Ersatz oder Ergänzung für das operativ erfasste Merkmal – ein Verfahren zu entwickeln, das die Arbeitsweise des Modells künftig regelmäßig überwacht und erneut trainiert.

# 1 Einleitung

Die BA berichtet mit der Statistik der gemeldeten Arbeitsstellen über Lage und Entwicklung des gegenwärtigen Arbeitskräftebedarfs am deutschen Arbeitsmarkt. Eine Meldepflicht für zu besetzende Stellen besteht für den Arbeitgeber nicht. Nur wenn die Arbeitgeber explizit die BA mit der Vermittlung geeigneter Arbeitsuchender beauftragen, finden diese Stellenangebote Eingang in die Statistik der gemeldeten Arbeitsstellen. Die erforderlichen Merkmale werden als Sekundärstatistik aus Prozessdaten der BA in Form einer Vollerhebung gewonnen. Definitionsgemäß handelt es sich dabei um einen Teilbereich des tatsächlichen gesamtwirtschaftlichen Stellenangebots.

Mit der Statistik der begonnenen Beschäftigungsverhältnisse aus der Beschäftigungsstatistik der BA[1] und der Stellenerhebung des Instituts für Arbeitsmarkt- und Berufsforschung (IAB)[2] liegen zwei weitere Datenquellen für Deutschland vor, die Sachverhalte in diesem Kontext erheben und berichten: Die Statistik der begonnenen Beschäftigungsverhältnisse berichtet über die realisierte Arbeitskräftenachfrage. Die IAB-Stellenerhebung gibt einen Einblick in die gesamtwirtschaftlichen Such- und Besetzungsvorgänge.

Somit können vergleichbare Merkmale dieser drei unterschiedlichen Datenquellen auf ihre übergreifende Kohärenz gegenübergestellt werden; s. Tabelle 1.

*Tabelle 1: Befristungsanteile sozialversicherungspflichtiger Einstellungen in %*

| Jahr | Statistik der gemeldeten Arbeitsstellen: Stellenabgang durch Besetzung | Beschäftigungsstatistik: begonnene sozialversicherungs-pflichtige Beschäftigungs-verhältnisse | IAB-Stellenerhebung: sozialversicherungs-pflichtige Neueinstellungen |
|---|---|---|---|
| 2020 | 16,4 | 40,1 | 34 |
| 2019 | 17,1 | 40,7 | 32 |
| 2018 | 19,7 | 42,9 | 36 |

*Quelle: Statistik der Bundesagentur für Arbeit; IAB*[3]*.*

Dabei stimmen die drei Datenquellen in der tendenziellen Abnahme des Befristungsanteils von 2018 bis 2020 annähernd überein; nur die IAB-Stellenerhebung signalisiert 2020 eine leichte Zunahme. Die Statistik der gemeldeten Arbeitsstellen weist durchgängig einen sehr niedrigen Befristungsanteil im Vergleich mit der Statistik der begonnenen Beschäftigungsverhältnisse und der IAB-Stellenerhebung aus, was das Risiko einer systematischen Unterschätzung birgt.

Der in der Statistik der gemeldeten Arbeitsstellen berichtete Befristungsstatus von Stellen stammt aus einem entsprechenden Eingabefeld im operativen Verfahren der BA, dem Vermittlungs-, Beratungs- und

[1] https://statistik.arbeitsagentur.de/DE/Navigation/Statistiken/Fachstatistiken/Beschaeftigung/Beschaeftigung-Nav.html

[2] https://www.iab.de/de/befragungen/stellenangebot.aspx (zuletzt abgerufen am 16.08.2021)

[3] Bundesagentur für Arbeit 2021 und Institut für Arbeitsmarkt- und Berufsforschung 2020.

Informationssystem (VerBIS). Dieses Feld repräsentiert die bei dem Kontakt mit den Arbeitgebern benannte Befristungseigenschaft. Unabhängig davon sind regelmäßig aber auch in den Stellenanzeigentexten der Arbeitgeber Informationen zur vorgesehenen Befristung enthalten, der sich direkt an potenzielle Arbeitsuchende richtet. Stellenanzeigentexte sind bisher nicht für die Datengenese in der Statistik der gemeldeten Arbeitsstellen nutzbar, da entsprechende Methoden bisher nicht verfügbar sind.

Die Statistik der BA erprobt nun erstmals Methoden des Text Mining für die Analyse von Stellenanzeigentexten. Der vorliegende Methodenbericht stellt die angewandte Methodik und die erzielten Ergebnisse vor, mit der systematisch die Validität des operativ erhobenen Merkmals „Befristung“ untersucht wurde.

Zunächst gibt Abschnitt 2 einen Überblick über die Statistik der gemeldeten Arbeitsstellen, die eine wichtige Komponente der Datenbasis für das Text Mining darstellt und die nicht-textbasierten Elemente für die Analysen liefert. Abschnitt 3 führt allgemein in Text Mining ein, bevor Abschnitt 4 die Methoden vorstellt, die speziell für den Themenkomplex „Befristungen“ angewendet werden: Welche Datenaufbereitungsschritte sind erfolgt, wie wurden die Texte quantifiziert und welche Modelle wurden getestet, um zu der Aussage zu gelangen, ob ein Stellenangebot auf Basis seines Textes als befristet oder unbefristet einzustufen ist? Zu welchen Ergebnissen gelangt das Modell? Abschnitt 5 zeigt auf, dass der Einsatz des Modells keine datenethischen Risiken birgt, und gibt einen Ausblick, wie die gewonnenen Erkenntnisse in die statistische Verarbeitung miteinbezogen werden können.

## 2 Die Statistik der gemeldeten Arbeitsstellen

Die Statistik der gemeldeten Arbeitsstellen umfasst die Arbeitsstellen für den ersten Arbeitsmarkt, die den Arbeitsagenturen und Jobcentern von den Arbeitgebern zur Vermittlung gemeldet wurden. Aus den Geschäftsprozessdaten werden monatlich für den Stichtag und den Monatszeitraum die gemeldeten Arbeitsstellen erhoben, im zentralen statistischen DataWarehouse (DWH) aufbereitet und anschließend publiziert.

Dabei stehen die gemeldeten Arbeitsstellen als Umfang der ungedeckten Arbeitskräftenachfrage der Zahl der Arbeitsuchenden als Abbild des Arbeitskräfteangebots gegenüber. Da jedoch keine Meldepflicht für zu besetzende Stellen besteht, kann es sich dabei immer nur um einen Teilbereich des vorhandenen gesamtwirtschaftlichen Stellenangebots handeln.

Ein Großteil des gesamtwirtschaftlichen Stellenangebots lässt sich durch die Statistik der gemeldeten Arbeitsstellen beschreiben, da die Betriebe einen Teil ihrer Stellenangebote – auch ohne bestehende Meldepflicht – den Agenturen und Jobcentern melden und zur Vermittlung freigeben.[4] Die Stellenangebote, die zur Vermittlung beauftragt werden, bilden die Erhebungseinheiten sowie die darin enthaltenen Stellen die Grundgesamtheit der Statistik.

[4] Im 4. Quartal 2018 lag der Anteil bei rd. 43 % (IAB-Stellenerhebung 2021).

Der Statistik der gemeldeten Arbeitsstellen liegt das Konzept eines Stock and Flow-Modells zugrunde.[5] Die Zugänge, Bestände und Abgänge bilden hierbei konsistente Messgrößen, die im zeitlichen Verlauf der dargestellten Beziehung folgen:

$$Anzahl\ Stellen_t = Anzahl\ Stellen_{t-1} + Zugang\ Stellen_t\ - Abgang\ Stellen_t$$ [6]

Die gemeldeten Arbeitsstellen umfassen sozialversicherungspflichtige, geringfügige und sonstige Arbeitsstellen.[7] Nicht Teil der Statistik der gemeldeten Arbeitsstellen sind die folgenden Beschäftigungsarten: kurzfristige Beschäftigungsmöglichkeiten bis sieben Kalendertage, Stellen für Freiberuflerinnen und Selbstständige[8], Stellen der Privaten Arbeitsvermittlung sowie geförderte Stellen des sogenannten „2. Arbeitsmarktes“ – im wesentlichen Arbeitsgelegenheiten nach § 16d SGB II.

Die wichtigsten Merkmale und Gliederungsdimensionen sind Sozialversicherungseigenschaft, Befristung, Arbeitszeit und -ort, Berufsbereich und Wirtschaftszweig. Neben der Differenzierung über Strukturmerkmale können die gemeldeten Arbeitsstellen anhand des erhobenen Arbeitsortes auch regional tief gegliedert ausgewiesen werden.

In Konzeption und Grundgesamtheit unterscheidet sich die Statistik der gemeldeten Arbeitsstellen von der Stellenerhebung des IAB, die ebenfalls eine Datenquelle zur ungedeckten Arbeitskräftenachfrage darstellt. Über eine repräsentative Betriebsbefragung bildet das IAB seit 2006 vierteljährlich die gesamtwirtschaftlichen Such- und Besetzungsvorgänge thematisch umfassender ab, als es der Statistik der BA möglich ist. Bei der Befragung werden auch Stellen erfasst, die von den Betrieben und Verwaltungen *nicht* an die Arbeitsagenturen und Jobcenter gemeldet werden. Die befragten Betriebs- und Verwaltungsstätten werden über eine Stichprobe ermittelt.

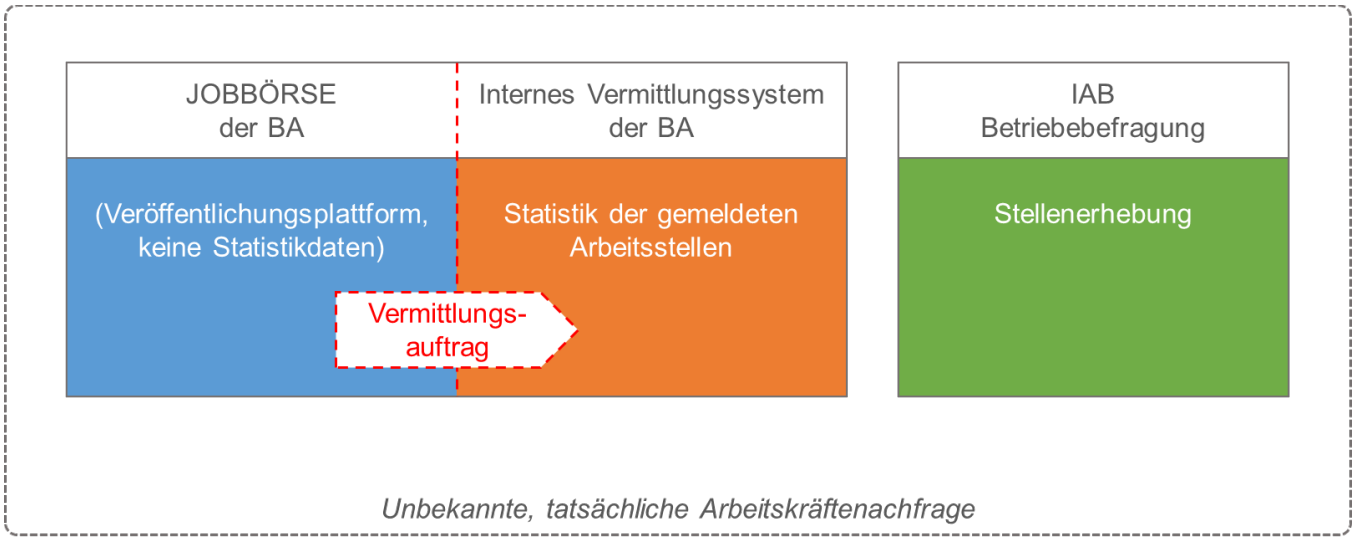
[5] Bundesagentur für Arbeit Statistik/Arbeitsmarktberichterstattung (2018)

[6] Werden keine weiteren Differenzierungen, wie bspw. nach der Art der Stelle vorgenommen, gilt diese Beziehung stets exakt für das gesamte Bundesgebiet. Zu beachten ist, dass sie für einzelne Gebietseinheiten oder andere Merkmale nur näherungsweise gilt, da Änderungen des Arbeitsortes bzw. anderer Merkmale und ihrer Ausprägungen nicht als Zu- und Abgänge gezählt werden.

[7] Zu sonstigen Arbeitsstellen gehören beispielsweise Stellen, die sozialversicherungsfrei sind, wie Praktikums- und Traineestellen, Beamte und Zeitsoldaten oder Stellen, bei denen die Sozialversicherungspflicht durch Arbeitgeber unspezifiziert ist und bspw. in Abhängigkeit der einzustellenden Arbeitsuchenden festgelegt werden soll.

[8] Diese Gruppen entsprechen nicht dem Begriff der Beschäftigung i. S. des § 7 SGB IV. Arbeitslose müssen Stellen, die mit einem unternehmerischen Risiko verbunden sind, nicht annehmen.

*Abbildung 1: Arbeitskräftenachfrage und Datenquellen*




*Quelle: Statistik der Bundesagentur für Arbeit*

Nach der Stellenerhebung des IAB waren im Jahr 2020 34 % der Neueinstellungen befristet; vgl. Tabelle 1. Der Befristungsanteil der begonnenen Beschäftigungsverhältnisse lt. Beschäftigungsstatistik lag 2020 bei 40 %. Die besetzten sozialversicherungspflichtigen Stellen der Statistik der gemeldeten Arbeitsstellen der BA hingegen zeigen im Jahresdurchschnitt 2020 mit 16 % einen deutlich geringeren Anteil an befristeten Arbeitsstellen.

Aufgrund der Befunde des IAB und der Beschäftigungsstatistik wäre in der Statistik der gemeldeten Arbeitsstellen ein höherer Befristungsanteil zu erwarten, als gemessen wird. Häufig enthalten die Stellenausschreibungstexte zusätzlich Hinweise zum Befristungsstatus. Einzelfallanalysen zeigen dann auch, dass die Informationen in den Stellentexten durchaus im Widerspruch zur operativen Befristungskennzeichnung stehen können. Herkömmliche statistische Methoden können bisher aus Texten keine Informationen gewinnen. Mit Hilfe von Sprachtechnologien, also durch Anwendung von Text-Mining-Methoden, lässt sich ein statistisches Modell zur Erkennung von befristeten Stellen entwickeln. Im Folgenden werden die Funktionsweise und die Einsatzmöglichkeiten von Text Mining näher erläutert und das Potenzial der Methode in Bezug zum vorliegenden Anwendungsfall gesetzt.

# 3 Text Mining

Nicht zuletzt durch die zunehmende Digitalisierung nimmt auch die Menge an verfügbaren digitalen Informationen in Form von natürlich-sprachlichen Texten, wie beispielsweise Freitextfelder in Befragungsbögen, Einträge auf Websites, stetig zu. Grundsätzlich ist es für entsprechend ausgebildete Menschen keine besondere Herausforderung, einer Textquelle gezielt Informationen zu entnehmen und diese zu strukturieren. Ein Problem stellt sich jedoch, wenn die große Menge an zu durchsuchenden Textquellen ein angemessenes Kontingent an Zeit und Kosten übersteigt. Dann kann es sinnvoll sein, diese Aufgabe an den Computer zu übertragen, der mit der Verarbeitung großer Datenmengen keine Schwierigkeiten hat. Die Herausforderung für den Computer bzw. für spezielle Programme hingegen ist es, die unterschiedlichen Repräsentationsformen und Formulierungen von Texten zu verstehen und diese entsprechend einzuordnen. Ein Leser mag diese Anforderung wie selbstverständlich umsetzen, aus maschineller Sicht jedoch ist es eine Hürde, die Inhalte sowie die Zusammenhänge von Textquellen zu erkennen und zu kategorisieren. Doch diese Hürde kann genommen werden: Prinzipiell ist es möglich, durch Text Mining sinnvoll Daten auszuwerten, die als natürlich-sprachliche Texte vorliegen.

Stellenanzeigentexte bieten ein interessantes Anwendungsbeispiel für die Informationsextraktion, da ein Korpus von Stellenanzeigentexten, der über mehrere Jahre zusammengestellt worden ist, wichtige Informationen zu den Entwicklungen am Arbeitsmarkt enthält.

Das IAB z. B. untersucht im Projekt „Stellenanzeigen im Lichte der ökologi-

**Grundzüge des Text Mining**

Text Mining ist eine interdisziplinäre Methode, die sich aus den Ergebnissen der Forschung zu Information Retrieval, Natural Language Processing (NLP oder maschinelle Sprachverarbeitung), Statistik, künstlicher Intelligenz, Machine Learning sowie Informationstheorie zusammensetzt (vgl. Hotho/Nürnberger/Paaß 2005).

Es lässt sich als auf Textdokumente bezogenes Data Mining beschreiben. Allerdings gibt es Unterschiede in Bezug auf die Datenaufbereitung, da auch linguistische Kriterien berücksichtigt werden müssen, um eine fehlende Datenstruktur rekonstruieren zu können (vgl. Rajman/Vesely 2004). Text Mining ermöglicht also die Verwendung von Textquellen unter Einbezug von deren Kontext und nicht nur die Auszählung von Schlüsselwörtern.

Wichtig dabei ist, dass Wörter innerhalb ihres Kontextes erkannt und verstanden werden müssen, um die gegebenen Zusammenhänge von Sätzen bzw. Inhalten sinnvoll analysieren zu können, da Worte in bestimmten Kontexten ihre Bedeutung verändern können. Man denke an das Beispiel „Mutter“, das in familiären Kontexten eine gänzlich andere Bedeutung als im handwerklichen Kontext hat, oder an Übersetzungen aus Fremdsprachen, die erst dann gelingen, wenn man den Kontext kennt: „You shall know a word by the company it keeps“ (Firth 1957: 11). Um diese Hürde zu nehmen, müssen statistische Methoden mit (Computer-)Linguistik verbunden werden.

Mit dieser Problemstellung befasst sich der Bereich des NLP, welche sich der Methoden der Informationsextraktion (Information Extraction, IE) bedient. In freien Texten sollen entsprechende Informationen erkannt und einheitlich strukturiert werden. So werden durch IE aus heterogenen Informationen, die in Form natürlicher Sprache vorliegen, strukturierte Daten, beispielsweise in Form von einheitlichen Datenbankeinträgen. Vorteilhaft daran ist, dass Daten, die auf diese Weise strukturiert worden sind, zur maschinellen Weiterverarbeitung und letzten Endes für statistische Auswertungen oder Data Mining Analysen genutzt werden können (vgl. Geduldig 2017).

schen und digitalen Transformation – Anpassungsprozesse in der beruflichen und regionalen Arbeitsnachfrage“ auf Basis von Online-Stellenanzeigen, wie sich die berufliche und regionale Arbeitsnachfrage vor dem Hintergrund der digitalen und ökologischen Transformation verändert.[9]

Methoden des Text Mining werden im Bereich der Arbeitskräftenachfrage bislang mittels so genanntem Web Scraping auf Stellenanzeigen in Online-Stellenbörsen in einem experimentellen Rahmen angewendet. Das Web Scraping ist eine Technik, bei der die Daten auf Internetseiten gezielt und automatisiert identifiziert, heruntergeladen und weiterverarbeitet werden (vgl. Rengers 2018).

Das hier beschriebene Vorhaben erprobt Text Mining erstmals anhand einer breiten Basis amtlicher Daten. Verwendet wird dieselbe Datenquelle, aus der auch die Statistik der gemeldeten Arbeitsstellen erstellt wird, angereichert um die Ausschreibungstexte der Stellenangebote. Diese werden bislang nicht für die statistische Berichterstattung der BA genutzt. Amtliche Statistikdaten als Grundlage für Text Mining heranzuziehen, hat mehrere Vorteile gegenüber Web Scraping mit Online-Stellenbörsen. So hat die Statistik der BA bei Text Mining mit gemeldeten Arbeitsstellen zum einen profunde Kenntnis über Struktur und Historizität der Datenquelle und zum anderen sind alle Merkmale verfügbar, die auch in den amtlichen Statistikdaten enthalten sind – und nicht nur Texte.

Zur grundlegenden Funktionsweise von Text Mining gehört die Transformation eines Textes in einen numerischen Vektor, auf dessen Grundlage ein Algorithmus den Zusammenhang zwischen einzelnen Merkmalen (Worten) und dem Klassenlabel („befristet“ oder „unbefristet“) erlernt. Damit dieser Lernprozess fachlich fundiert verläuft, kodierten Statistik-Fachkräfte die erstmaligen Vorhersagen des Modells. Das heißt, durch die Statistik-Fachkräfte wurde überprüft, ob das Modell bei der erstmaligen Vorhersage einen Stellenangebotstext richtig als befristet (oder unbefristet) erkannt hat. Als befristet haben die Statistik-Fachkräfte solche Stellenangebote klassifiziert, deren Text entweder den ausdrücklichen Hinweis „befristet“ bzw. „Befristung“ enthielt oder aus denen anderweitig aus dem Text hervorging, dass es sich um ein befristetes Arbeitsverhältnis handelt (z. B. um eine Elternzeitvertretung). War dies nicht der Fall, galten die Stellenangebote als unbefristet. Widersprüchliche Fälle („befristet/unbefristet“) wurden nicht in den Datensatz mit den Basisdaten (vgl. 4.1) aufgenommen.

Seitens der Statistik-Fachkräfte besteht keine interessengeleitete Präferenz für die eine oder andere Klassifikationsentscheidung. Diese wurden zudem im Team intersubjektiv abgeglichen. Durch dieses Vorgehen kann ein Bias, der im Rahmen der manuellen Kodierung entstanden sein könnte, weitgehend ausgeschlossen bzw. als so geringfügig eingeschätzt werden, dass die Methode auch den Anforderungen der Datenethik entspricht.

Die auf diese Weise von den Statistik-Fachkräften kodierten Datensätze stellten den Datensatz dar, mit dem das Modell dann seine Vorhersagen erneut berechnete. Mit der nun fachlich fundierten Kodierung der Datensätze als befristet oder unbefristet konnte es auf verbesserter Grundlage erlernen, welche strukturellen Eigenschaften eines Stellenausschreibungstextes mit den Klassen „befristet“ und „unbefristet“ einhergehen. Das Modell versuchte anhand der so erkannten Zusammenhänge, möglichst präzise die Kodierungen der Statistik-Fachkräfte durch die eigenen Vorhersagen zu treffen. Durch mehrmaliges

[9] https://www.iab.de/138/section.aspx/Projektdetails/k200701805 (zuletzt abgerufen am 16.08.2021)

Wiederholen dieses Prozesses wurden die Vorhersagen des Modells sukzessive verbessert. Das Modell wurde also trainiert (vgl. Abb. 2 in Abschnitt 4.3).

Auf Basis der kodierten Daten wurden verschiedene im Text Mining übliche Algorithmen getestet, um bestmögliche Voraussagen hinsichtlich des Befristungsstatus von Stellenangeboten zu treffen: der Random Forest, die Logistische Regression, die Support Vector Machine, XGBoost-Klassifikator und Neuronale Netze; s. a. Abschnitt 4.3.

# 4 Methoden

Dieser Abschnitt beschreibt den Ansatz und das Vorgehen, anhand derer der Befristungsstatus von Stellen ermittelt wurde. Dem Befristungsmodell liegt ein Klassifikationsmodell zugrunde, das automatisiert erkennen soll, ob ein Stellenbeschreibungstext Hinweise auf den Befristungsstatus der Stelle liefert. Testläufe im Vorfeld haben gezeigt, dass entscheidend für die Qualität der Modellvorhersage die zugrundeliegenden Trainingsdaten sind. Dafür wurde von Statistik-Fachkräften der BA eine manuelle Kodierung der Daten vorgenommen.

Im Folgenden werden der Prozess der Datenaufbereitung, das verwendete Machine Learning Modell und die Testergebnisse beschrieben (vgl. Abschnitt 4.3).[10]

## 4.1 Basisdaten

Für diesen experimentellen Ansatz des Text Minings wurden die ursprünglichen Stellenangebotsdaten um die Stellenanzeigentexte angereichert und dem operativ erfassten Befristungsmerkmal gegenübergestellt. Als Basisdaten diente ein Datensatz, der 3.306 Stellenangebote umfasst. Die Stellenangebote wurden zufällig als Stichprobe aus dem Bestand des Berichtsmonats Oktober 2018 gezogen. Dieser Datensatz wurde von den Statistik-Fachkräften kodiert. Die Verteilung der unbefristeten und befristeten Stellenanzeigen im Ursprungsdatensatz ist der nachfolgenden Tabelle 2 zu entnehmen:

*Tabelle 2: Häufigkeitsverteilung der Zielvariable Befristung*

| Befristungsmerkmal | Ausprägung | | | Summe |
|---|---|---|---|---|
| | Ja | Nein | Arbeitnehmer-überlassung | |
| | 1 | 2 | 3 | 4 |
| abgeleitet aus Stellenanzeigentext | 1.274 | 1.823 | 209 | 3.306 |
| operativ erfasst | 450 | 2.856 | - | 3.306 |

*Quelle: Statistik der Bundesagentur für Arbeit.*

Somit gehen 1.274 Stellenangebote (38,5 %) mit Hinweisen auf eine Befristung im Stellenanzeigentext und 1.823 Stellenangebote (55,1 %) ohne Hinweise in den Datensatz ein. Stellenangebote aus der Arbeitnehmerüberlassung enthalten oft keinen eindeutigen Hinweis auf eine Befristung, aber auf eine Übernahmeoption, sodass die Befristungseigenschaft nicht sicher festgestellt werden kann. Diese 209 Stellenangebote sind daher nicht Teil des Basisdatensatzes.

Für die Entwicklung des Klassifikationsmodells wird der manuell kodierte Basisdatensatz in eine Trainings- und eine Testmenge geteilt. Im Trainingsdatensatz sind 2.473 Stellenangebote, also circa 80 % der verfügbaren Fälle, enthalten. Diese werden für Entwicklung und Analyse – das sogenannte Training

[10] Dem Team „Advanced Analytics“ der BA, insbes. Alexander Rebmann und Joachim Seitz, sei gedankt, das mit methodischer Expertise Datenaufbereitung, Datenanalyse und deren Dokumentation erst ermöglicht hat.

– des Modells genutzt. Im Testdatensatz sind mit 624 Stellenangeboten etwa 20 % der Fälle enthalten, die für die Evaluation des Modells verwendet werden.

## 4.2 Textaufbereitung

Damit ein Klassifikationsmodell trainiert werden kann, müssen die Textdaten zunächst in mehreren Aufbereitungsschritten verarbeitet werden. An deren Ende steht die Transformation des Textes in einen numerischen Vektor (Vektorisierung),[11] auf dessen Grundlage der Algorithmus den Zusammenhang zwischen den Merkmalen (Worten) und dem Klassenlabel („befristet“ oder „unbefristet“) erlernt. Im Folgenden werden kurz die Schritte erläutert, die beim Training für das Modell mit der besten Vorhersage angewandt wurden.

Im ersten Datenaufbereitungsschritt werden sämtliche Datumsangaben durch das Wort „Datum“ ersetzt. Zudem wurden Sonderzeichen und sogenannte Stoppwörter, also Begriffe, die sehr häufig auftreten und gewöhnlich keine inhaltliche Relevanz für die Dokumente haben und somit für die Klassifikationsentscheidung überflüssig und irreführend sind (z. B. „der“, „die“, „das“), aus dem Datensatz entfernt.

Als Merkmale, die für die Vektorisierung herangezogen werden, wurden N-Gramme auf Buchstabenebene erzeugt. Das heißt, der Text wird in eine Menge von Zeichenfolgen zerlegt, die n benachbarte Buchstaben umfassen. Das Merkmal kann die Ausprägungen 0 und 1 annehmen. Das Merkmal nimmt den Wert 0 an, wenn es nicht im Text vorhanden ist, und den Wert 1, wenn es im Text vorhanden ist.

## 4.3 Modelltraining und Evaluation

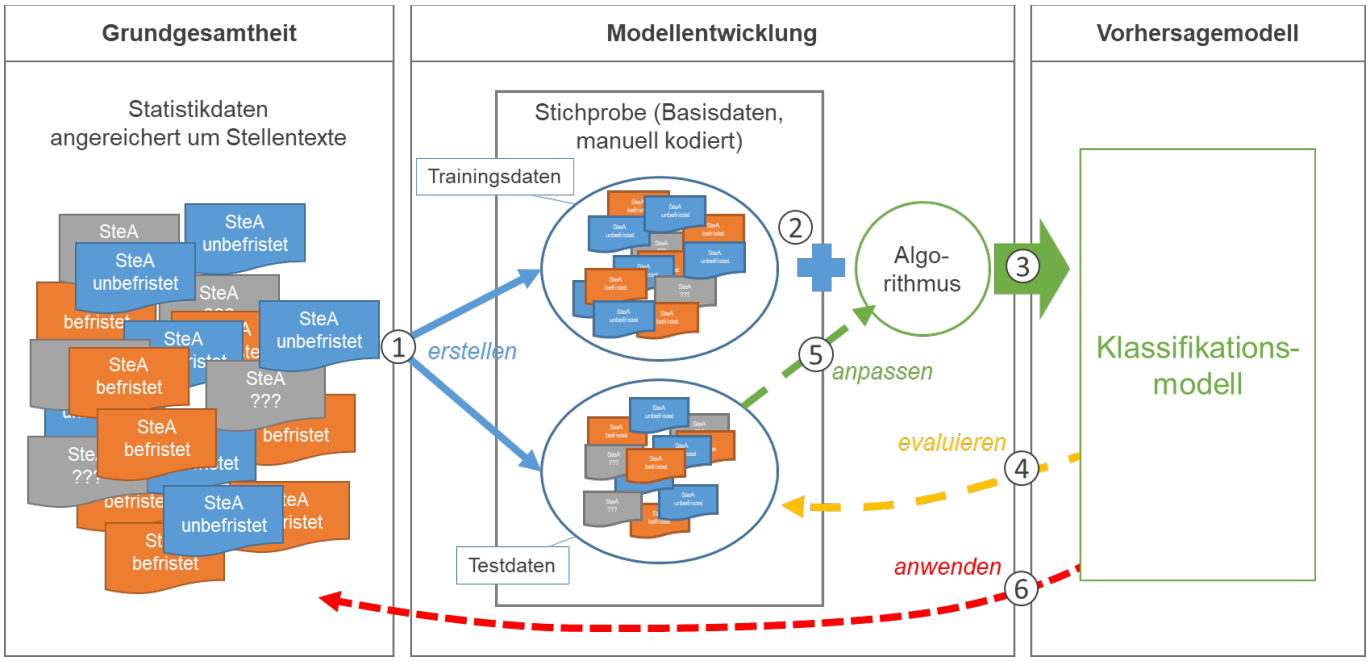
Um eine möglichst gute Modellanpassung zu erreichen, wurden während der Modellentwicklung fünf unterschiedliche, im Text Mining gängige Algorithmen hinsichtlich ihrer Güte getestet:[12]

- Random Forest
- Logistische Regression
- Support Vector Machine
- XGBoost-Klassifikator
- Neuronale Netze

[11] Durch Vektorisierung werden Worte als Zahlen dargestellt und in einem mehrdimensionalen Raum verortet. Bedeutungsähnliche Worte bzw. Worte, die oft im selben Kontext vorkommen, liegen dort nah beieinander: „Die Worte Eiche, Ulme und Birke könnten sich in einer Ecke zusammenfassen, während Krieg, Konflikt und Streit in einer anderen zusammenfließen.“ (Katzlberger 2021).

[12] Vgl. z. B. https://www.intellspot.com/text-mining-algorithms/ (zuletzt abgerufen am 16.08.2021)
https://www.kaggle.com/diveki/classification-with-nlp-xgboost-and-pipelines (zuletzt abgerufen am 16.08.2021)

*Abbildung 2: Modellentwicklung*




*Quelle: in Anlehnung an: [https://blog.frankfurt-school.de/machine-learning-modern-data-analytics-artificial-intelligence/](https://blog.frankfurt-school.de/machine-learning-modern-data-analytics-artificial-intelligence/) (zuletzt abgerufen am 16.08.2021)*

Die Modelle für den Trainingsdatensatz wurden mit diesen fünf Algorithmen trainiert und ihre Genauigkeit anhand des Testdatensatzes evaluiert. Aus der Parameteroptimierung auf den Basisdaten geht der XGBoost-Klassifikator als das Modell hervor, das mit seinen Vorhersagen die Klassifizierung im Trainingsdatensatz am genauesten trifft.

Das Neuronale Netz und der XGBoost-Klassifikator erzielen bei der Verwendung der Testdaten dieselbe Genauigkeit (Accuracy)[13], während der Random Forest eine geringere Genauigkeit aufweist, vgl. Tabelle 3. Dabei gilt: Je näher die Werte an 1 liegen, desto genauer ist die Modellvorhersage. Da der XGBoost-Klassifikator und das Neuronale Netz die gleiche Modellgüte in Bezug auf „Befristung Precision“[14] und „Befristung Recall“[15] aufweisen, der XGBoost-Klassifikator aber leichter zu interpretieren ist, liegt der Fokus auf diesem Modell.[16]

[13] Accuracy ist der Anteil aller richtigen Vorhersagen des Modells (true positives und true negatives) an allen seinen Vorhersagen.

[14] Precision bezeichnet den Anteil der richtig vorhergesagten Befristungen (true positives) an den insgesamt als Befristungen vorhergesagten Stellenangeboten (precision = True Positives / (True Positives + False Positives).

[15] Recall bezeichnet den Anteil der richtig vorhergesagten Befristungen (true positives) an den insgesamten Vorhersagen des Modells (sowohl true als auch false positives).

[16] Unter dem Gesichtspunkt der Datenethik ist dieses Modell vorzuziehen, weil es nachvollziehbarer und verständlicher und somit eine Auseinandersetzung damit möglich ist.

*Tabelle 3: Vergleich der drei besten Modellergebnisse*

| *Modell* | Metrik | | |
|---|---|---|---|
| | Accuracy | Befristung Precision | Befristung Recall |
| | 1 | 2 | 3 |
| *Neuronales Netz* | 0,957 | 0,968 | 0,926 |
| *XGBoost Klassifikator* | 0,957 | 0,975 | 0,919 |
| *Random Forest* | 0,880 | 0,856 | 0,853 |

*Quelle: Statistik der Bundesagentur für Arbeit*

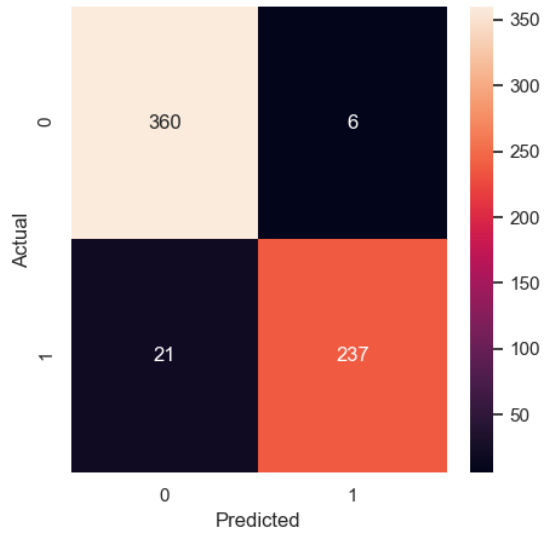
Das XGBoost-Modell sagt für 95,7 % der Stellenbeschreibungstexte korrekt voraus, ob Hinweise auf eine Befristung enthalten sind oder nicht (Accuracy = 0,957). Es gehört zu den Ensemble Methoden – genauer zu den Gradient Boosting Methoden.[17] Das Modell setzt sich aus einer Vielzahl von Entscheidungsbäumen zusammen. Der Trainingsprozess erfolgt sequentiell. Die Entscheidungsbäume werden nach und nach dem Modell hinzugefügt. Das Ziel der nachfolgenden Bäume ist, die Fehlklassifikationen vorheriger Bäume auszugleichen. Dabei implementiert der XGBoost-Algorithmus das Gradient Boosting mit effektiven Regularisierungstechniken und ist auf Effizienz für große Datenmengen optimiert.[18]

Interessant sind nun Fälle, bei denen die Stellenbeschreibung einen Befristungshinweis enthält. Diesen erkennt der XGBoost-Klassifikator zu etwa 91,9 % richtig. Die Sicherheit der Vorhersage (Precision) liegt bei 97,5 %. Die Kennzahlen können aus der Konfusionsmatrix (Abb. 3) abgeleitet werden.

[17] Gradient Boosting ist eine maschinelle Lerntechnik für Regression, Klassifizierung und andere Aufgaben, die ein Vorhersagemodell in Form eines Ensembles von schwachen Vorhersagemodellen, typischerweise Entscheidungsbäumen, erzeugt.

[18] Vgl. https://xgboost.readthedocs.io/en/latest/tutorials/model.html (zuletzt abgerufen am 16.08.2021)

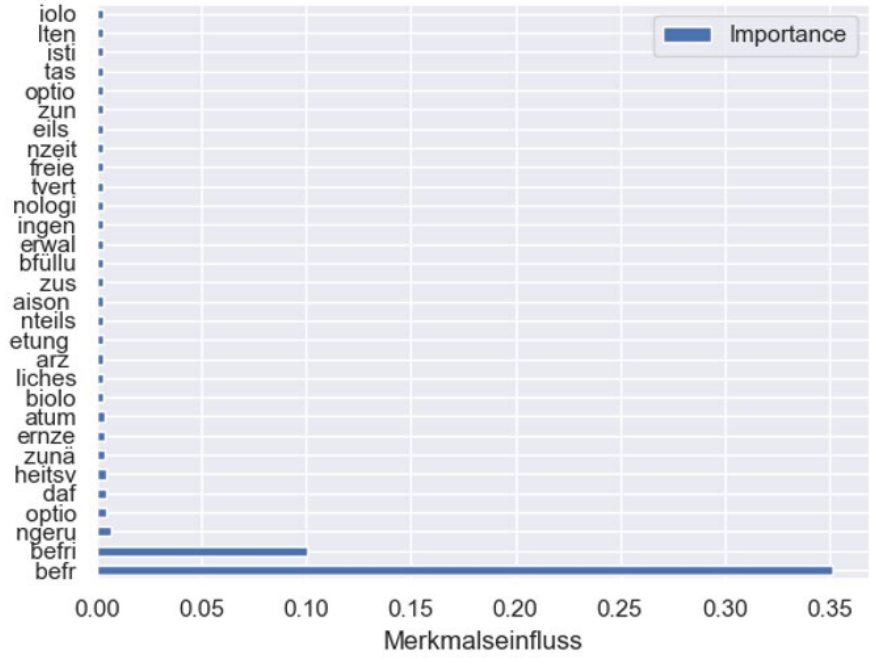
*Abbildung 3: Konfusionsmatrix*




*Quelle: Statistik der Bundesagentur für Arbeit*

Die Konfusionsmatrix stellt die tatsächlichen den abgeleiteten Zielvariablen (Befristung Text) gegenüber. Als Lesebeispiel: Das untere linke Rechteck in Abbildung 3 zeigt die Zahl der vom Modell als nicht befristet vorhergesagten (predicted), aber durch Statistik-Fachkräfte als befristet markierten (actual) Stellenanzeigen. Entsprechend sind im oberen rechten Eck die als befristet vorhergesagten, aber faktisch unbefristet markierten Stellenanzeigen enthalten.

*Abbildung 4: Bedeutende Merkmale bei der Klassifikationsentscheidung*




*Quelle: Statistik der Bundesagentur für Arbeit*

Anhand der Merkmalsbedeutung kann ermittelt werden, welche Merkmale aus dem Textkorpus für die Klassifikationsentscheidung ausschlaggebend waren (siehe Abbildung 4). Die mit Abstand wichtigsten Merkmale sind Buchstabenfolgen des Wortes „befristet“ („befr“: 35 %, „befri“: 10 %). Aber auch andere eindeutige Merkmale haben Einfluss auf die Klassifikationsentscheidung, wie Buchstabenfolgen aus den Wörtern Saison, Option, Verlängerung, Vertretung, Elternzeit und Datum.

Aufgrund der Ergebnisse, ihrer guten Interpretierbarkeit und der hohen Modellgüte fiel die fachliche Entscheidung auf den XGBoost-Klassifikator. Das Modell könnte die statistische Information, ob ein Stellenangebot befristet oder unbefristet ausgeschrieben ist, gegenüber dem jetzigen Stand verbessern. Deshalb geht das Modell in einen Testbetrieb, während dem es auf die Statistikdaten angewendet wird; s. u. Abschnitt 5.2.

# 5 Risikoprüfung und Ausblick

## 5.1 Risikoprüfung „Rechtsrahmen für Künstliche Intelligenz“ der EU

Die hier vorgestellte Methode basiert auf Algorithmen, welche die Statistik der BA bisher nicht angewendet hat. Ob es sich dabei bereits um eine Art künstlicher Intelligenz handelt, sei dahingestellt. Jedenfalls soll ein Programm Texte untersuchen und damit eigenständig Fragen beantworten, die bisher lediglich Menschen beantworten können. Deshalb ist eine zumindest kursorische Überprüfung i.S. des von der Europäischen Kommission vorgeschlagenen Rechtsrahmens für künstliche Intelligenz[19] angezeigt.

1) Das Ziel der vorliegend angewendeten Algorithmen ist eine statistische Klassifizierung von Stellenangeboten unter dem Kriterium der Befristung des Stellenangebots.
   - Damit ist keine operative Anwendung verbunden; die Ergebnisse werden operativ nicht wirksam (Rückspielverbot aus der Statistik).
   - Individualrechte sind nicht berührt, da hier lediglich Arbeitsstellen bzw. Texte hierüber, nicht Menschen oder menschliches Verhalten beurteilt werden.
   - Die Methode wirkt nicht auf den Einzelfall, denn sie dient einer statistischen Aussage über die Häufigkeitsverteilung von Befristungsangaben in einer Grundgesamtheit von Arbeitsstellen.
   - Das Risiko für eine „wirksame Durchsetzung der bestehenden Rechtsvorschriften im Bereich der Sicherheit und der Grundrechte“ (Europäische Kommission 2021) durch den Einsatz der hier beschriebenen Methodik erscheint somit als sehr gering.

2) Weitere Prüfungen der Methodik bieten sich an bei der Stichprobenauswahl, bei der Beurteilung der Stichprobenfälle, bei der Entscheidung für einen der zur Auswahl stehenden Algorithmen sowie bei den Kriterien, die der eingesetzte Algorithmus verwendet:
   - Stichprobenauswahl: Die Stellenangebote, die in die Stichprobe einflossen, wurden unter Beachtung einer angemessenen Wirtschaftszweigstruktur, ansonsten rein zufällig ausgewählt. Eine Diskriminierung ist deshalb ausgeschlossen.
   - Die Klassifizierung der Stichprobenfälle bezüglich Befristung haben mehrere Statistik-Fachkräfte vorgenommen. Bei menschlichen Entscheidungen lassen sich zwar Verzerrungen nicht völlig ausschließen; sie sind jedoch vorliegend angesichts des Themas Befristung und auch dadurch minimiert, dass nicht nur eine einzelne Person beteiligt war.
   - Die Entscheidung zwischen den beiden erfolgreichsten Algorithmen fiel zugunsten desjenigen aus, dessen Ergebnisse leichter erklärbar und damit nachvollziehbar sind. Undurchsichtigkeit ist somit reduziert und die Ergebnisse sind hinterfragbar.

[19] Vgl. Europäische Kommission (2021)

- Unter den für die Klassifizierung bedeutenden Merkmalen, s. Abbildung 4, finden sich keine Begriffe oder Begriffsteile, die die Möglichkeit einer Diskriminierung in sich bergen.

Die eingesetzte Methode bzw. der eingesetzte Algorithmus ist also grundsätzlich risikoarm. Weitere Prüfungen weisen die hohe Sicherheit beim Einsatz der Methode aus.

### 5.2 Ausblick: Einsatz im DataWarehouse der Statistik der BA

Um Erfahrungen mit den modellgenerierten Daten zu gewinnen und bei Eignung eine Integration in die reguläre, statistische Datengenese im DataWarehouse vorzubereiten, wird das Befristungsmodell derzeit einem einjährigen, manuellen Testbetrieb unterzogen. Der Analyseprozess der Stellentexte erfolgt dabei noch nicht vollautomatisiert. Die um die Stellentexte erweiterte Datenbasis des jeweils aktuellen Berichtsmonats soll mit Hilfe des binären Befristungsmodells „Befristung Ja/Nein“ ausgewertet werden.

Sobald Ergebnisse hierzu vorliegen, müssen sie validiert und auch auf ihre Kombinierbarkeit mit anderen Auswertungsmerkmalen, wie Zeit, Region, Wirtschaftszweig usw., überprüft werden. Besonders wichtig wird der Vergleich mit dem operativ erfassten Merkmal „Befristung“ sein.

Die Erprobungsphase dient auch der technischen Weiterentwicklung der bisherigen statistischen Datengenese, insbesondere der Einbettung des Befristungsmodells in die automatisierten Abläufe und Abhängigkeiten.

Neben der Entscheidung über eine statistische Veröffentlichung als Ersatz für das operativ erfasste Merkmal oder als Ergänzung zu diesem Merkmal ist außerdem ein Verfahren zu entwickeln, das die Arbeitsweise des Modells regelmäßig überwacht und erneut trainiert.

## Literaturverzeichnis

Bundesagentur für Arbeit (2021): Sozialversicherungspflichtige Beschäftigungsverhältnisse (Quartalszahlen), Nürnberg.

Bundesagentur für Arbeit Statistik/Arbeitsmarktberichterstattung (2018): Grundlagen: Qualitätsbericht, Statistik der gemeldeten Arbeitsstellen, Nürnberg.

Europäische Kommission (2021): Neue Vorschriften für künstliche Intelligenz – Fragen und Antworten (https://ec.europa.eu/commission/presscorner/detail/de/QANDA_21_1683, zuletzt abgerufen am 13.07.2021).

Firth (1957): A Synopsis of Linguistic Theory, 1930-1955. In: Studies in Linguistic Analysis. Oxford.

Geduldig (2017): Muster und Musterbildungsverfahren für domänenspezifische Informationsextraktion. Ein Bootstrapping-Ansatz zur Extraktion von Kompetenzen aus Stellenanzeigen, Uni Köln.

Hohendanner (2018): Reform der befristeten Beschäftigung im Koalitionsvertrag: Reichweite, Risiken und Alternativen, IAB-Kurzbericht, 16/2018, Nürnberg.

Hotho/Nürnberger/Paaß (2005): A Brief Survey of Text Mining. In: Gesellschaft für Linguistische Datenverarbeitung e. V. (Hrsg.): Zeitschrift für Computerlinguistik und Sprachtechnologie. Band 20, Heft 1.

Institut für Arbeitsmarkt- und Berufsforschung (2020): Befristungen bei Neueinstellungen, Nürnberg.

Institut für Arbeitsmarkt- und Berufsforschung (2021): Befristete Beschäftigung in Deutschland 2020, Nürnberg.

Institut für Arbeitsmarkt- und Berufsforschung (2021): IAB-Stellenerhebung, Nürnberg. (https://www.iab.de/de/befragungen/stellenangebot/aktuelle-ergebnisse.aspx, zuletzt abgerufen am 21.06.2021)

Katzlberger, Michael (2021): Word2Vec – verständlich erklärt. (https://katzlberger.ai/2019/06/25/word2vec-verstaendlich-erklaert/, zuletzt abgerufen am 26.07.2021)

Lemke/Wiedemann (2016): Text Mining in den Sozialwissenschaften. Grundlagen und Anwendung zwischen qualitativer und quantitativer Diskursanalyse, Wiesbaden.

Rajman/Vesely (2004): From Text to Knowledge: Document Processing and Visualization: a Text Mining Approach. In: Sirmakessis (Hrsg.): Text Mining and its Applications. Studies in Fuzziness and Soft Computing, Berlin, Heidelberg.

Rengers (2018): Internetgestützte Erfassung offener Stellen. Machbarkeitsstudie im Rahmen eines ESS-net-Projekts zu Big Data. In: Statistisches Bundesamt: WISTA – Wirtschaft und Statistik, 5/2018, Wiesbaden.

Roßbach (2017): Machine Learning, Modern Data Analytics and Artificial Intelligence – What's new? (https://blog.frankfurt-school.de/machine-learning-modern-data-analytics-artificial-intelligence/, zuletzt abgerufen am 21.06.2021).

Statistik der Bundesagentur für Arbeit (2018): Grundlagen: Methodenbericht - Befristete Beschäftigung, Methodische Hintergründe und Ergebnisse, Nürnberg.

## Statistik-Infoseite

Im Internet stehen statistische Informationen unterteilt nach folgenden Themenbereichen zur Verfügung:

- Arbeitsmarkt und Grundsicherung im Überblick
- Arbeitslose, Unterbeschäftigung und Arbeitsstellen
- Ausbildungsstellenmarkt
- Beschäftigung
- Förderung und berufliche Rehabilitation
- Grundsicherung für Arbeitsuchende (SGB II)
- Leistungen SGB III
- Berufe
- Bildung
- Daten zu den Eingliederungsbilanzen
- Einnahmen/Ausgaben
- Familien und Kinder
- Frauen und Männer
- Langzeitarbeitslosigkeit
- Migration
- Regionale Mobilität
- Wirtschaftszweige
- Zeitreihen
- Amtliche Nachrichten der BA
- Kreisdaten

Die Methodischen Hinweise der Statistik bieten ergänzende Informationen.

Das Glossar enthält Erläuterungen zu allen statistisch relevanten Begriffen, die in den verschiedenen Produkten der Statistik der BA Verwendung finden.

Abkürzungen und Zeichen, die in den Produkten der Statistik der Bundesagentur für Arbeit vorkommen, werden im Abkürzungsverzeichnis bzw. der Zeichenerklärung der Statistik der BA erläutert.