Kanton Zürich
Staatskanzlei und Statistisches Amt

Einsatz Künstlicher Intelligenz in der Verwaltung

Bericht zu einer Pilot-Anwendung der Studie zu rechtlichen und ethischen Fragen

Herausgeber
Kanton Zürich
Staatskanzlei und Statistisches Amt

Autorinnen und Autoren
Christian Ruiz Palmero, Statistisches Amt
Laure Stadler, Statistisches Amt
Franziska Moser, Staatskanzlei
Cornelia Wodnik, Staatskanzlei

Mai 2023, Version 2.0

Inhalt

1 Management Summary

In einem Pilotversuch setzt die Verwaltung die im Bericht zum «Einsatz Künstlicher Intelligenz in der Verwaltung: rechtliche und ethische Fragen» vorgestellten Massnahmen und Checklisten, die den Kanton Zürich beim Einsatz von KI unterstützen sollen, erstmals ein und verankert das Wissen um Einsatzmöglichkeiten für KI-Anwendungen in der kantonalen Verwaltung, was deren Akzeptanz unter den Mitarbeitenden erhöhen soll. Das dient dazu, weitere Anwendungsbereiche zu finden und den Aufwand einer KI-Lösung einschätzen zu können.

Der Bereich Digitale Verwaltung der Staatskanzlei wählte in Zusammenarbeit mit dem Statistischen Amt für diesen Pilotversuch den verwaltungsinternen Prozess der Triage von Vorstössen aus dem Kantonsrat und setzte auf ein Entscheid unterstützendes System. In der Umsetzung wurden zwei KI-basierte Ansätze auf sehr unterschiedliche Datengrundlagen angewendet:

- In einem klassischen Machine-Learning-Ansatz wurde ein Algorithmus mit Daten bereits triagierter Geschäfte auf Direktions- und Amtsebene trainiert. Der Algorithmus lernt gestützt auf bereits gefällte Entscheide das selbstständige Triagieren nach demselben Muster. Auf Direktionsebene war die Datenquantität und -qualität hoch. Auf Amtsebene wird die Zuteilung der Geschäfte noch nicht lange, anders und von Direktion zu Direktion unterschiedlich dokumentiert, weswegen die Datenqualität und -quantität deutlich niedriger waren.
- In einem experimentelleren Ansatz hat ein Algorithmus die semantische Nähe der parlamentarischen Vorstösse zu den Aufgaben der Direktionen und Ämter beurteilt.

Die erreichten Ergebnisse dieser spezifischen Triage-Anwendung zeigen, dass der Einsatz von Machine Learning als Entscheidungsunterstützung in der Verwaltung unter bestimmten Rahmenbedingungen möglich ist. Die befragten Verwaltungsmitarbeitenden begrüssten den Einsatz von Künstlicher Intelligenz zur Entlastung und Qualitätssteigerung in der Verwaltungsarbeit und erkannten während des Projektverlaufs weitere potenzielle Einsatzmöglichkeiten.

Die Ziele des Pilotprojekts wurden somit erreicht. Die gewonnenen Erkenntnisse und Erfahrungen zeigen, dass die Rahmenbedingungen geschaffen bzw. verbessert werden müssen, um die Potenziale möglichst gut auszunützen. Dabei gibt es nicht nur Handlungsbedarf in diesem konkreten Anwendungsfall, sondern auch in der generellen richtigen Weichenstellung zur Erreichung von Innovation und Prozessverbesserung. Der Handlungsbedarf für den konkreten Anwendungsfall ist überschaubar und kann gezielt adressiert werden. Der generelle Handlungsbedarf mit den vorgeschlagenen Massnahmen richtet sich nach den Themenbereichen der Leitsätze «gemeinsam digital unterwegs»[1]: Organisation, Daten, Infrastruktur, Recht und Leistungen.

[1] Siehe RRB Nr. 1362/2021

2 Ausgangslage

Im April 2018 setze der Regierungsrat die Strategie Digitale Verwaltung 2018–2023 fest. Um die strategischen Ziele zu erreichen, hat der Kanton in einem Vorprojekt den Einsatz von Künstlicher Intelligenz in der Verwaltung untersuchen lassen. Das Ergebnis wurde im Februar 2021 im Bericht «Einsatz Künstlicher Intelligenz in der Verwaltung: rechtliche und ethische Fragen» veröffentlicht. Der Bericht beleuchtet umfassend die Möglichkeiten der Verwendung von KI in der kantonalen Verwaltung und stellt Checklisten zur Verfügung, mit denen KI-Vorhaben auf ethische Konflikte geprüft werden können. Verfasst wurde der Bericht durch das Team der Juristischen Fakultät der Universität Basel mit Unterstützung durch AlgorithmWatch Schweiz.

Im Nachgang zur Veröffentlichung der Studie hat die Staatskanzlei in Zusammenarbeit mit dem Statistischen Amt beschlossen, nach einer geeigneten KI-Anwendungsmöglichkeit zu suchen, die im Rahmen eines Pilotprojekts und unter Berücksichtigung der im Vorprojekt erarbeiteten Checklisten umgesetzt werden kann. Das Projektteam hat sich auf einen Anwendungsfall geeinigt, der einerseits einen konkreten Nutzen innerhalb der Verwaltungstätigkeit schafft und anderseits aus Sicht der betroffenen Daten im Sinne des Datenschutzes unkritisch ist: die automatisierte Triage von parlamentarischen Vorstössen innerhalb der kantonalen Verwaltung.

2.1 Ziele des Pilotprojekts

Mit dem Pilot-Anwendungsfall wollen die Staatskanzlei und das Statistische Amt praktische Erfahrungen mit Künstlicher Intelligenz bzw. Machine Learning innerhalb der Verwaltung sammeln. Ziel ist ein Erfahrungsgewinn sowohl auf breiter Ebene als auch für die mit der Umsetzung betrauten Stellen der Staatskanzlei und des Statistischen Amtes.

Durch die Mitarbeit in der Entwicklung und den späteren Einsatz dieser ersten Anwendung werden Verwaltungsmitarbeitende mit dem neuen Ansatz vertraut gemacht. Es ist wichtig, aufzuzeigen, dass und welche Routineprozesse teilautomatisiert werden können. Wenn Mitarbeitende wissen, wie durch Teilautomatisierung Verwaltungsprozesse effizienter durchgeführt werden können, sind sie befähigt, weitere Einsatzmöglichkeiten zu erkennen und zu melden.

Die Auswertung des Pilotprojekts hilft der Staatskanzlei und dem Statistischen Amt, Aufwand und Nutzen für weitere Einsatzbereiche von KI-Algorithmen einzuschätzen. Dies ermöglicht eine realistische Planung weiterer KI-Anwendungsfälle. Im kleinen Rahmen des konkreten Anwendungsfalls soll der Mehrwert durch Machine Learning gegenüber einer manuellen Bearbeitung aufgezeigt werden.

2.2 Anwendungsfall des Pilotprojekts

Die Triage und Behandlung von parlamentarischen Vorstössen sichert eine gute Verbindung zwischen Parlament und Verwaltung: Der Kantonsrat übt die verfassungsgebende und die gesetzgebende Gewalt aus. Der Regierungsrat ist die oberste leitende und vollziehende Behörde des Kantons. Er wahrt die Verfassung und setzt die Gesetze, Verordnungen und Beschlüsse des Kantonsrates um. Mit parlamentarischen Vorstössen kann der Kantonsrat dem Regierungsrat Aufträge erteilen sowie Auskünfte oder Berichte verlangen.

Die Kantonsratsmitglieder stellen die parlamentarischen Vorstösse den Parlamentsdiensten elektronisch zu. Die Parlamentsdienste prüfen die Vorstösse, passen sie gegebenenfalls formell an und leiten sie elektronisch über das Geschäftsverwaltungssystems GEVER[2] an die Staatskanzlei weiter. Die Staatskanzlei teilt das Geschäft inhaltlich zu und sendet es zur weiteren Verteilung an das Generalsekretariat der zuständigen Direktion. Innerhalb der Direktionen nehmen

[2] In der Schweiz bekannt als «GEVER»; das Pendant in Deutschland ist die «E-Akte».

die Generalsekretariate die weitere Zuteilung an das zuständige Amt oder die zuständige Fachstelle vor. Ist ein Geschäft in der Zuständigkeit der Staatskanzlei, so entscheidet die Staatsschreiberin, an welchen Bereich das Geschäft weitergeleitet wird. Dies erfolgt durch die elektronische Weiterleitung vom GEVER des Regierungsrates an das GEVER der Staatskanzlei.

Wer welche Geschäfte behandelt, ist eine verantwortungsvolle Entscheidung. Die Zuteilung auf oberster Ebene (zuständige Direktion oder Staatskanzlei) erfolgt durch die Staatsschreiberin. Sie stützt sich bei ihrem Entscheid auf einen Vorschlag ihrer Assistenz oder der Assistenz des Rechtsdienstes. Dieser Vorschlag zur Zuteilung erfolgt aufgrund verschiedener Kriterien:

- Verzeichnis der Aufgaben der Direktionen gemäss Verordnung über die Organisation des Regierungsrates und der kantonalen Verwaltung (VOG RR)
- Suche nach ähnlichen Fällen aus der Vergangenheit, um die Wahrscheinlichkeit einer richtigen Zuteilung zu erhöhen
- Vorwissen und Erfahrung der Personen in diesen Rollen

Die Staatskanzlei bearbeitet wöchentlich 10–15 parlamentarische Vorstösse. Es handelt sich somit nicht um ein Massengeschäft, doch sind es regelmässig wiederkehrende Aufgaben, die personelle Mittel binden und von einer personenunabhängigen Konstanz profitieren. Mitarbeitenden, die für die Triage von parlamentarischen Vorstössen zuständig sind, soll durch eine KI-basierte Lösung eine Entscheidungsunterstützung geboten werden.

2.3 Rahmenbedingungen/Scope des Pilotprojekts

Ziel des Pilot-Anwendungsfalls war, einen Prototyp zu erstellen. Dieser läuft auf einer Umgebung, die nicht in die produktive Umgebung und Prozesse der betroffenen Personen eingebunden ist. Am Ende des Pilotprojekts ist zu entscheiden, ob der Prototyp zu einem Minimal Viable Product (MVP) ausgebaut werden soll. Dies würde bedeuten, dass der Prototyp in die Systemlandschaft der beiden Rollen innerhalb der Staatskanzlei und gegebenenfalls weiterer Stellen integriert wird. Dieser Entscheid ist durch die Auftraggeberin zu treffen, wobei die Umsetzung des MVP im Rahmen der strategischen Initiative Daten denkbar wäre.

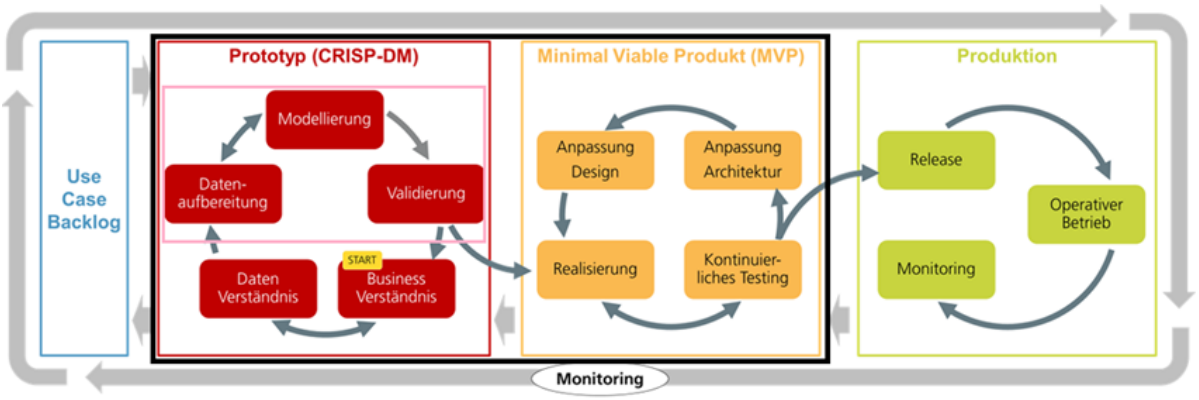

Abbildung 1: Scope (Bildquelle: Data Science Competence Center, Bundesamt für Statistik, 2021)

2.4 Rechtliche Situation und Schutzbedarf für das Pilotprojekt

Die rechtliche Situation wurde durch eine Juristin des Bereichs Digitale Verwaltung vor Start der Arbeiten geprüft.
Dabei wurde festgestellt, dass der KI-Pilot durch folgende, bereits bestehende Rechtsgrundlagen abgedeckt ist:

- Allgemeine Handlungsgrundsätze der Verwaltung: Art. 70 Abs. 2 der Kantonsverfassung (LS 101) und §§ 32 und 33 des Gesetzes über die Organisation des Regierungsrates und der kantonalen Verwaltung (LS 172.1)
- Statistikgesetz (LS 431.1) (§§ 2ff. u.a. betreffend Analyse und Interpretation von Daten mit statistischen Methoden)

Insbesondere wurde auch festgestellt, dass weder besondere Personendaten (siehe § 8 Abs. 2 Gesetz über die Information und den Datenschutz vom 12. Februar 2007 [IDG, LS 170.4]) bearbeitet noch Grundrechte gefährdet oder verfassungsmässige Rechte eingeschränkt werden.

2.5 Risiken des Pilotprojekts

Eine Analyse der Risiken des KI-Piloten erfolgte anhand der Triage-Checkliste für KI-Systeme, die innerhalb der Studie «Einsatz Künstlicher Intelligenz in der Verwaltung: rechtliche und ethische Fragen, Schlussbericht vom 28. Februar 2021 zum Vorprojekt IP6.4» erstellt wurde.
Dabei konnten folgende Kernaussagen erarbeitet werden:

- Der KI-Pilot arbeitet nur mit parlamentarischen Vorstössen, die keine Einzelpersonen betreffen. Insbesondere sind keine Verfügungen oder Entscheide betroffen, mit denen Rechte und Pflichten einhergehen.
- Es werden auch keine politischen Entscheidungen (z.B. Wahl oder Volksabstimmung) beeinflusst.
- Das technische System beruht nicht auf einem statistischen Modell des menschlichen Verhaltens oder der persönlichen Merkmale, sondern stellt ein Regelwerk anhand logischer Werte zur Verfügung. Es werden keine Risikoabschätzungen getroffen, nur Wahrscheinlichkeitsabschätzungen.
- Das grösste Risiko des KI-Piloten besteht darin, dass ein Vorstoss aufgrund eines nicht korrekten oder missinterpretierten Vorschlages an eine nicht zuständige Direktion geroutet wird. Diese falsche Entscheidung kann aber jederzeit manuell korrigiert werden.
- Der «Schaden» ist somit vollständig reversibel. Es kommt höchstens zu einem vertretbaren Zeitverlust bei der Bearbeitung des parlamentarischen Vorstosses.

Anhand dieser Kernaussagen und der Auswertung der Checkliste kommen wir zum Schluss, dass es sich um eine risikoarme Anwendung handelt.

Zusätzlich prüften wir mithilfe des Flussdiagramms der «Checkliste Transparenzbericht», ob ein solcher Bericht zu erstellen ist. Dieser würde aufzeigen, ob die ethischen Richtlinien zum Einsatz KI-basierter Systeme, wo notwendig, eingehalten werden.

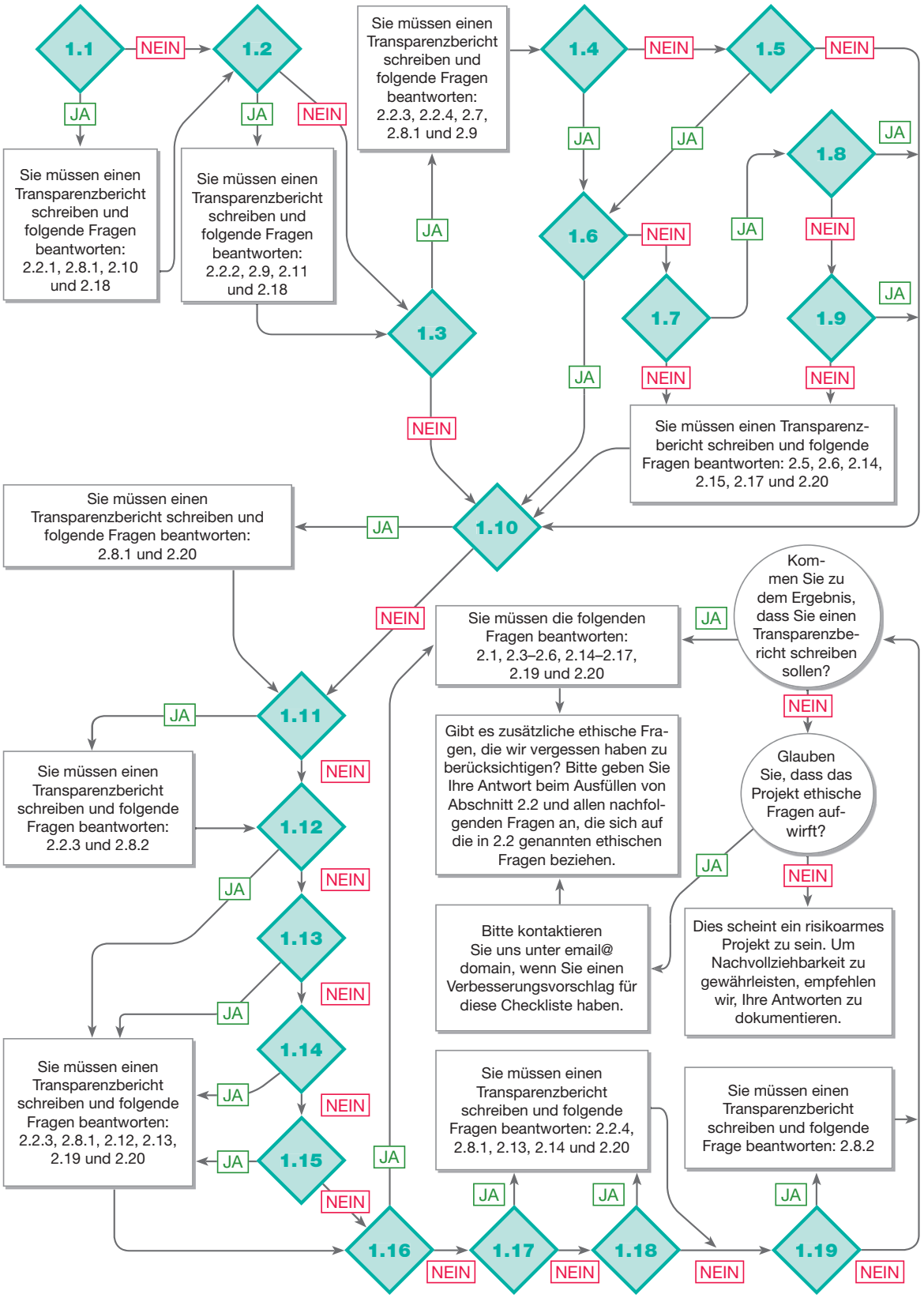

Abbildung 2: Übersicht Flussdiagramm «Checkliste Transparenzbericht»

Die Auswertung des Flussdiagramms ergab, dass kein Transparenzbericht notwendig ist.

3 Realisierung des Pilotprojekts

Um von der Idee zum Prototyp zu gelangen, mussten in der Realisierung folgende Schritte durchgeführt werden:

- Interviews mit Nutzenden, um den Prozess zu verstehen und den Bedarf herauszufiltern
- Auswertung der Interviews und Definition der Anforderungen
- Entscheid über das technische Vorgehen
- Umsetzung und Testen des Prototyps

Diese werden in den folgenden Kapiteln weiter ausgeführt.

3.1 Prozesserhebung durch Interviews

Damit die betroffenen Personen tatsächlich durch den Einsatz von Künstlicher Intelligenz entlastet werden können, mussten ihr tatsächlicher Bedarf und ihre Anforderungen an eine KI-Lösung aufgenommen werden. Dazu wurden im September 2021 fünf Interviews mit den potenziellen Nutzenden der Staatskanzlei und der Direktion der Justiz und des Innern geführt. Wir haben uns auf diese beiden Bereiche konzentriert, da wir hier am schnellsten und einfachsten an die Nutzenden herantreten konnten. Auch sind die notwendigen Arbeitsmittel (GEVER, manuelle Schritte) sehr ähnlich.
Wir verwendeten einen einheitlichen Fragenkatalog, damit die Antworten miteinander vergleichbar waren.

Aufgrund der Interviews lässt sich der Prozess der Triage parlamentarischer Vorstösse wie folgt beispielhaft nachzeichnen:

1. Die Staatskanzlei erhält einen parlamentarischen Vorstoss aus dem Kantonsrat mittels GEVER.
2. Die Staatskanzlei bestimmt, welche Direktion für die Behandlung dieses Vorstosses hauptsächlich verantwortlich ist (Triage 1).
   → Für Vorstösse, die nicht die Direktion der Justiz und des Innern betreffen, endet der Prozess hier.
3. Innerhalb der Direktion der Justiz und des Innern erfolgt eine weitere Triage: Das Geschäft wird in einem nächsten Schritt den Aufgabenbereichen zugeordnet (Triage 2).
4. Innerhalb eines Aufgabenbereichs wird in einem letzten Schritt das Geschäft dem zuständigen Amt zugeordnet (Triage 3) und überstellt.

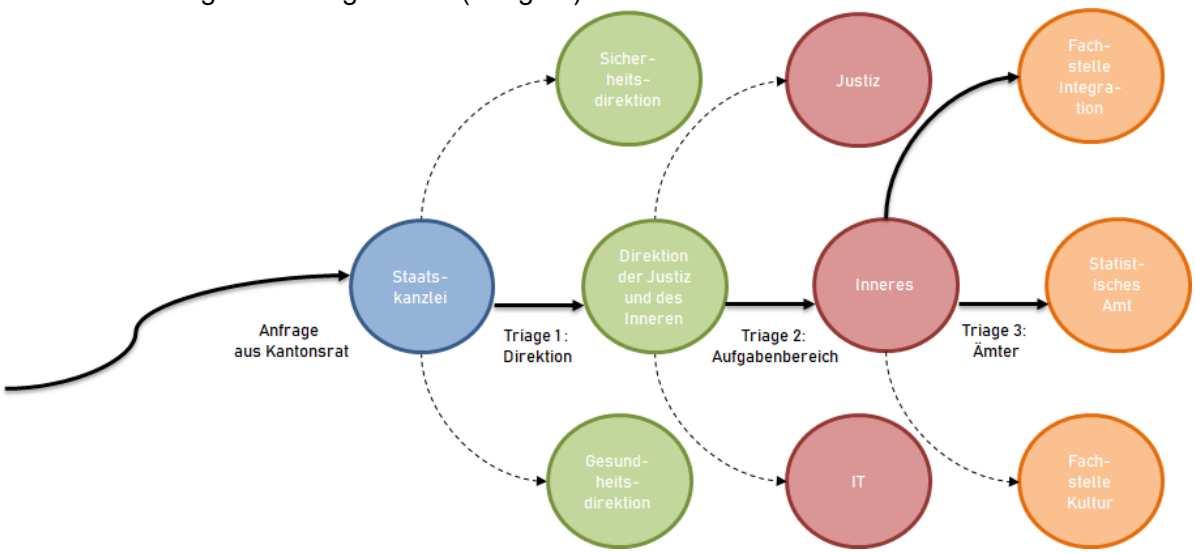

Abbildung 3: Prozessdarstellung der drei Triagen der Direktion der Justiz und des Innern

Während der Interviews wurden noch weitere Themen angesprochen, die für einen späteren Ausbau der Triagefunktion relevant sein können:
- Es wäre hilfreich, wenn bei der Übermittlung der Vorstösse der Prozess weiter automatisiert und somit die manuellen Eingriffe reduziert werden könnten.
- Den Interviewten war wichtig, dass das System korrigiert werden kann und in der Lage ist dazuzulernen.
- Die Triage muss nachvollziehbar sein → ändern sich Zuständigkeiten oder Organisationsstrukturen, muss das dem System mitgeteilt werden können.

3.2 Technische Umsetzung

Die in den Interviews identifizierten Bedürfnisse an die automatisierte Triage von Kantonsratsgeschäften haben wir in einem System zur Entscheidungshilfe umgesetzt. Dazu haben wir zwei verschiedene, sich ergänzende Ansätze kombiniert.
Zunächst werden die beiden Ansätze vorgestellt. Dabei unterscheiden sich im ersten Ansatz die organisatorischen Ebenen (Direktionsebene [1a]) und Amtsebene [1b]) hinsichtlich Prozess und Datengrundlage deutlich. Danach wird in diesem Kapitel die Kombination der beiden Ansätze abgebildet. Das Kapitel endet mit einigen Schlussbemerkungen zur angewendeten Methodik und zur Infrastruktur.

3.2.1 Erster Ansatz basierend auf ML-Training bestehender Daten

Direktionsebene
Der erste Ansatz fokussiert sich auf ein klassisches Machine-Learning-(ML-)Training anhand bestehender bzw. vergangener Daten. Da wir wissen, welche Direktionen und Ämter sich mit den erledigten Vorstössen aus dem Kantonsrat beschäftigt haben, können wir den Algorithmus mit diesen alten Daten trainieren, sodass eine Vorhersage für einen neuen Datenpunkt getroffen werden kann. Das ist ein klassischer Ansatz, da man sowohl abhängige (zu erklärende) Variablen (zuständige Direktion oder Amt) als auch unabhängige Variablen (die Texte der Geschäfte) hat. Die erstellten Modelle basieren dabei auf vortrainierten Transformer-Modellen (DistilBERT) und weiteren Grundlagen im Bereich des sogenannten Natural Language Processing (NLP), um die Texte in nutzbare Vektoren zu transformieren.

Der Algorithmus auf Direktionsebene ist die «Haupt-Triage» dieses Pilotprojekts: Die erste Triage zwischen den Direktionen findet in der Staatskanzlei statt und leitet somit die Vorstösse vom Kantonsrat direkt an die zuständige Direktion weiter. Dort werden die Vorstösse wiederum an das zuständige Amt weitergeleitet.

Die Datenquelle der Geschäfte des Kantonsrates ist als offene Behördendaten (OGD) frei verfügbar. Man kann alle Geschäfte seit 1987 abrufen.[3]

Amtsebene(n)
Im Weiteren haben wir ML-Modelle innerhalb der Direktionen trainiert, um nachzubilden, wie die Triagen auf Amtsebene funktionieren. Das Generalsekretariat der jeweiligen Direktion empfängt das Geschäft von der Staatskanzlei und leitet dieses dann an das zuständige Amt und die zuständigen Personen weiter. Dabei sehen die direktionsinternen, oben beschriebenen Prozesse in der befragten Direktion der Justiz und des Innern sowie der Staatskanzlei unterschiedlich aus und müssen auch dementsprechend in der Erstellung der Algorithmen berücksichtigt werden.

[3] Es sind leider nicht alle Dokumente seit 1987 in einem verwendbaren Format vorhanden. Der Text vieler älterer Dokumente lässt sich nicht automatisiert extrahieren.

Die Datengrundlage auf Amtsebene ist eine andere als auf Direktionsebene. Zwar verwenden wir ebenfalls die offenen Behördendaten der Geschäfte, müssen diese aber mit den Daten des direktionsinternen GEVER verknüpfen. Letzteres speichert die historisierten Informationen zu den Geschäften einschliesslich der damit verbundenen Aufgaben, die von Ämtern und Personen ausgeführt wurden.

Die Datenlage ist hier deutlich schlechter. Das GEVER wurde erst vor einigen Jahren eingeführt, daher ist die Datenquantität allgemein gering. Auch die Datenqualität ist nicht ideal. Im Fall der Direktion der Justiz und des Innern konnten die Geschäfte anhand der Kantonsratsnummer verknüpft werden. Im Fall der Staatskanzlei war dies nicht möglich, da nicht dieselben Nummern verwendet wurden. In beiden Fällen war ein grösseres Data Cleaning notwendig, um die Daten für ein ML-Training nutzbar zu machen (z.B. Korrigieren von Schreibfehlern in den erfassten Titeln o.Ä.).

Da die Daten für die Staatskanzlei nicht einfach verknüpft werden konnten, war das Vorgehen dort nochmals anders: Wir haben, aufgrund der fehlenden Verknüpfung mit den vollständigen Texten der Geschäfte, nur die im GEVER hinterlegten Beschreibungen bzw. Titel der Geschäfte verwendet. Wir wollten somit zumindest ausprobieren, ob dies eine einfache Zwischenlösung wäre. Eine korrekte Verknüpfung wäre über Umwege oder manuell zwar möglich, aber umständlich gewesen. Da die Datenmenge in der Staatskanzlei besonders gering war, zeigte sich, dass es sich dieser Aufwand nicht lohnt.

Hingegen war die Anzahl der Geschäfte seitens der Direktion der Justiz und des Innern grösser als bei der Staatskanzlei, und wir konnten die Geschäfte über die Kantonsratsnummer verknüpfen. Aber auch hier war die Datenmenge der Geschäfte nicht ausreichend.

Ein weiterer Unterschied zwischen den beiden Teilnehmenden ist die zweistufige Triage bei der Direktion der Justiz und des Innern. Wie Abbildung 3 zeigt, triagiert das Generalsekretariat dort die Geschäfte zunächst in fünf bzw. vereinfacht gesagt in drei Themengruppen: Gesellschaft, Recht, Anderes (es gibt hier sehr wenig Fälle in den Kategorien IT, HR und Controlling, weshalb diese hier in einer Gruppe zusammengefasst werden). Bei den zuständigen Personen findet dann die eigentliche Triage an das zuständige Amt statt. Das haben wir in unseren Versuchen ebenfalls berücksichtigt und zwei verschiedene Algorithmen trainiert: ein Algorithmus, der in die drei Themengruppen einteilt, und einer der dann Vorschläge für das zuständige Amt macht.

3.2.2 Zweiter Ansatz basierend auf semantischer Nähe

Ein zweiter Ansatz kann ergänzend zum ersten verwendet werden, um auch ohne vergangene Entscheidungen eine Einschätzung zu treffen, die eine zusätzliche Information liefert. Dabei wird die semantische Nähe zwischen einem Text und einer vordefinierten Zuständigkeit ermittelt. Diese semantische Nähe wird anhand von Zuständigkeiten ermittelt, die in Anhang 1 VOG RR (LS 172.11) aufgelistet sind.

Hierzu wird die semantische Nähe zwischen dem Geschäft und den einzelnen aufgelisteten Zuständigkeiten berechnet. Bei diesem Ansatz wurde eine Methode verwendet, die sprachlich bereits auf deutsche Sprache vortrainiert wurde.

172.11 VOG RR

Anhang 1: Zuständigkeitsbereiche der Direktionen
(§ 58)

A. Direktion der Justiz und des Innern[14, 16, 17, 19, 27]

1. Justizvollzug einschliesslich Begnadigungen
2. Strafverfolgung Erwachsene einschliesslich Rechtshilfe und Auslieferungen
3. Jugendstrafrechtspflege
4. Filmwesen
5. Gemeindewesen einschliesslich Finanz- und Lastenausgleich
6. Bezirkswesen
7.[41] Zivilstands-, Bürgerrechts- sowie Meldewesen und Einwohnerregister
8. Handelsregister
9. Statistik
10. Archivwesen
11. Berufliche Vorsorge und Stiftungsaufsicht
12. Opferhilfe bei Straftaten
13. Kulturförderung
14. Gleichstellung von Frau und Mann
15. Integrationsfragen

Abbildung 4: Auszug Anhang 1 VOG RR

Der Vorteil ist, dass kein Machine-Learning-Modell von uns trainiert werden muss. Deshalb spricht man bei diesem Algorithmus auch von «zero-shot learning»[4].

Der Nachteil ist, dass dem Algorithmus keine weiteren Zusammenhänge bekannt sind.
Um ein Beispiel zu geben: Die VOG RR sieht eine Unterscheidung zwischen genereller Statistik und spezifischer Bildungsstatistik vor. Die generelle Statistik gehört in den Bereich der Direktion der Justiz und des Innern, die spezifische Bildungsstatistik dagegen zur Bildungsdirektion. Das sind spezifische Informationen, die ein Algorithmus, der nur die semantische Nähe betrachtet, nicht berücksichtigen kann. Aber auch über dieses Beispiel hinaus, ist es selbstverständlich, dass ohne Training eines spezifischen Kontextes die Genauigkeit der Aussage bei einem solchen Algorithmus sehr schlecht ist. Nichtsdestotrotz könnten die Angaben eine zusätzliche Hilfe für Mitarbeitende sein, sofern diese die Empfehlungen nicht ohne Reflexion akzeptieren, sondern wissen, dass der Algorithmus nur die semantische Nähe ausdrückt. Zusätzlich muss man auch erwähnen, dass die vortrainierten Algorithmen in Deutsch gegenwärtig noch unpräziser sind als die in Englisch.

[4] Siehe z.B. en.wikipedia.org/wiki/Zero-shot_learning

3.2.3 Kombination der beiden Ansätze

Diese beiden Ansätze können grundsätzlich kombiniert verwendet werden und ergänzen sich. Als Beispiel für eine Kombination wird in Abbildung 5 das Resultat der beiden Algorithmen zur Anfrage KR-Nr. 168/2021 betreffend Maskentragen aus epidemiologischer Sicht gezeigt. Auf der linken Seite (Resultat 1) ist die Berechnung aus dem klassischen ML-Ansatz sichtbar. Der Algorithmus errechnet aufgrund von früheren Geschäften eine Wahrscheinlichkeit von 99%, dass für dieses Geschäft die Gesundheitsdirektion zuständig ist. Auf der rechten Seite (Resultat 2) ist das Ergebnis des zweiten Ansatzes aufgrund semantischer Nähe sichtbar. Das Element «Gesundheitswesen, Epidemiewesen ... und Rettungswesen», das hier durch den Algorithmus mit einer engen semantischen Nähe errechnet wurde, ist in der VOG RR der Gesundheitsdirektion zugeordnet. In diesem Fall passen somit die Resultate beider Algorithmen mit der tatsächlich zuständigen Direktion zusammen. Im zweiten Algorithmus sind auch andere Elemente wie «Staatsrechtliche Massnahmen im Bereich der nationalen und internationalen Aussenbeziehungen» und «Schutz vor Naturgefahren» als relevant eingestuft worden. Dies kann in diesem Kontext auch Sinn ergeben, wobei die Einschätzung der oder des Mitarbeitenden hier sehr zentral ist.

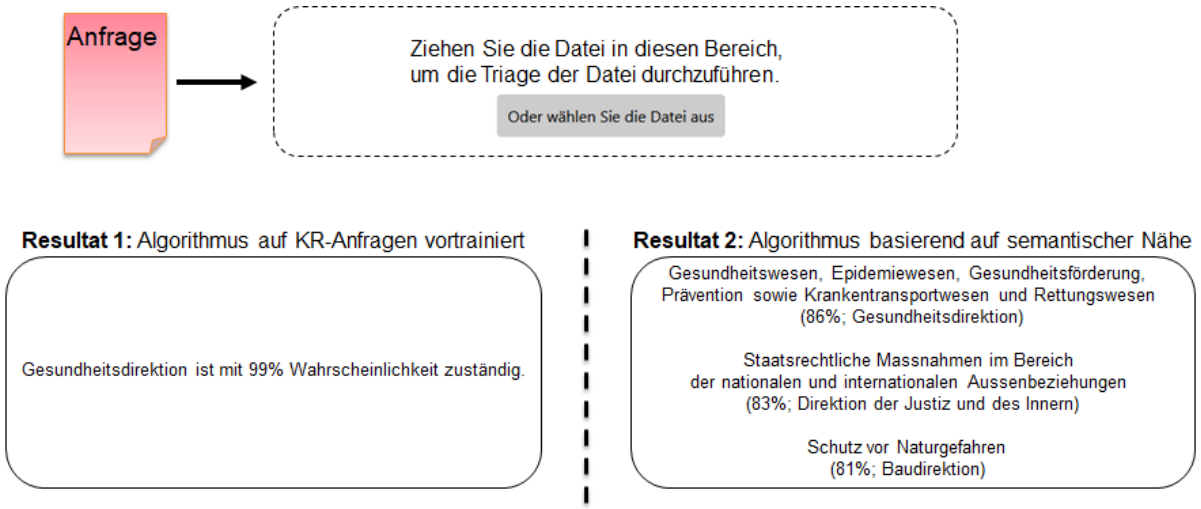

Abbildung 5: Beispiel eines Resultats zweier ergänzender Algorithmen

3.2.4 Verwendete Machine-Learning-Methodik

Die Machine-Learning-Modelle wurden methodisch nach neustem Stand der Technik trainiert:

- Die Daten wurden strikt in Trainings- und Testdaten getrennt (80% und 20%).
- Es wurde darauf geachtet, dass die zu berechnenden Klassen mittels Upsampling ausgeglichen sind.
- Es wurde eine sigmoide Aktivierungsfunktion verwendet, damit mehrere Klassen pro Geschäft «richtig» sein können. Die Resultate wurden auch Berechnungen mit Softmax gegenübergestellt.
- Es wurden sowohl die Indikatoren Loss als auch Accuracy berechnet, wobei auch hier die Resultate bei den Validierungsdaten für die Beurteilung ausschlaggebend waren.
- Es wurde nicht nur die gesamte Accuracy, sondern auch die Resultate der einzelnen Klassen geprüft, um zu überprüfen, ob die Algorithmen über alle Klassen hinweg möglichst sinnvolle Resultate liefern.
- Um die besten Resultate zu erreichen, wurden mehrere Parameter wie Epochen, Lernraten und Batchgrössen variiert.
- Es wurden für beide Ansätze vortrainierte Transformer-Modelle auf Deutsch verwendet (jeweils DistilBERT und ein auf GBERT basierter Algorithmus).

- Data Cleaning musste an verschiedenen Stellen im Prozess angewendet werden, wodurch die Anzahl der verfügbaren Daten schwankt.
- Die errechneten Klassen entsprechen je nach zugrunde gelegten Daten den zuständigen Direktionen, Ämtern oder Abteilungen.

3.2.5 Verwendete Infrastruktur

Es wurde keine fortgeschrittene Infrastruktur verwendet, sondern mit bereits vorhandenen Ressourcen gearbeitet. Die Infrastruktur blieb somit bescheiden, aber konnte dennoch zur Zielerreichung des Pilotprojekts beitragen. Die Modelle wurden innerhalb der Infrastruktur des Statistischen Amtes trainiert. Für das Pilotprojekt wurde ein Webserver im Intranet aufgesetzt. Auf einer einfachen Oberfläche (wie in der Abbildung 5) konnte man eine Datei auswählen oder in den Browser ziehen und sich dann die Auswertungen anzeigen lassen. Auf dem Server waren die vortrainierten Modelle installiert und konnten mit den neuen Datenvektoren ausgeführt werden.

3.3 Ergebnisse

3.3.1 Ergebnisse für Triage auf Direktionsebene

Bei der Triage auf Direktionsebene wurde eine sehr hohe und brauchbare Accuracy erreicht. Damit die Vorhersage der verschiedenen Direktionen im Datensatz ausgeglichener sind, wurde ein Upsampling verwendet. Damit wurde eine Test-Accuracy von 92,5% erreicht. Ohne Upsampling wurden sogar 93,44% erreicht, was aber weniger repräsentativ ist.

Die folgende Tabelle zeigt die Genauigkeit der Vorhersage im Testset für die einzelnen Direktionen. Die Baudirektion wurde insgesamt am häufigsten falsch vorhergesagt. Jedoch konnten wir mit Upsampling den Wert von 87,3% auf 88,3% leicht erhöhen.

Direktion	Spezifische Accuracy ohne Upsampling	Häufigkeit im Datenset ohne Upsampling	Spezifische Accuracy mit Upsampling
Bildungsdirektion	93,0%	1303	92,9%
Sicherheitsdirektion	91,9%	820	91,8%
Baudirektion	87,3%	1802	88,3%
Direktion der Justiz und des Innern	93,4%	1007	92,3%
Finanzdirektion	92,6%	984	92,2%
Volkswirtschafts-direktion	89,0%	1587	89,2%
Gesundheits-direktion	95,4%	820	95,2%
Staatskanzlei	98,4%	147	98,4%

Tabelle 1: Resultate der Triage auf Direktionsebene

Insgesamt wurden hier 8470 Geschäfte behandelt, die aus der Geschäftsdatenbank des Kantonsrates extrahiert werden konnten. Wie bereits erwähnt, gibt es mehr PDF-Dateien in der Geschäftsdatenbank. Allerdings konnten viele insbesondere die ältesten Dokumente darin nicht automatisiert aus den PDF extrahiert werden. Die Tabelle zeigt, dass es für alle Direktionen ausser der Staatskanzlei eine grosse Anzahl von Geschäften gegeben hat. Mit nur 147 Geschäften ist die Staatskanzlei ein Outliner und die Accuracy von 98,4% muss somit auch mit Vorsicht betrachtet werden, da die Aussagekraft beschränkt ist.

Wir haben die Resultate einiger Geschäfte manuell überprüft und sind ebenfalls auf gute Resultate gekommen. Wenn der Algorithmus eine andere Direktion vorschlägt, konnte man häufig nachvollziehen, wieso dieses Resultat zustande kam. Das vorhandene Wissen, die Erfahrung des Verwaltungskontextes und die inhaltlichen Fragestellungen sind zentral, um aus solchen «falschen» Vorschlägen als Mitarbeitende in der Praxis dennoch einen Mehrwert in Form von Informationen zu ziehen, mit denen man dann inhaltlich begründet für eine andere Direktion entscheidet. Beispielsweise wenn es um Behörden geht, die womöglich in einem Thema nicht offiziell die Federführung haben, aber im Prozessverlauf sicher involviert werden.

Es gibt gute Einwände, die wir leider nicht berücksichtigen konnten. Beispielsweise haben sich die Zuständigkeiten im Laufe der Zeit verändert. War in den 1990er-Jahren noch Direktion A für Thema X zuständig, wurden die Zuständigkeiten beispielsweise in den letzten Jahren der Direktion B übertragen. Solche Änderungen berücksichtigt das Training des Algorithmus nicht, da die einzigen Variablen der Text des Geschäfts und die dafür zuständige Direktion waren. Alle anderen Faktoren werden im Training ignoriert. Man könnte den Algorithmus verbessern, indem man beispielsweise die Daten rückwirkend korrigiert – also so, dass im Trainingsdatensatz die heute zuständigen Direktionen statt der damals zuständigen Direktionen aufgeführt werden. Eine Lösung für eine automatisierte rückblickende Zuordnung haben wir im Rahmen dieses Pilotprojekts vorläufig nicht gefunden.

3.3.2 Ergebnisse für Triage auf Amtsebene

3.3.2.1 Direktion der Justiz und des Innern

Die Interviews mit den Mitarbeitenden haben ergeben, dass die Triage in zwei Schritte ausgeführt wird: Zunächst werden die Geschäfte vom Generalsekretariat entweder in Recht, Gesellschaft oder Anderes[5] triagiert. Danach wird in einem zweiten Schritt das Geschäft den einzelnen Ämtern (und Direktionen) zugeordnet, da hier mehr inhaltliches Fachwissen und Erfahrung benötigt wird. Deshalb haben wir versucht, diesen zweistufigen Prozess auch mit Algorithmen abzubilden.

Schritt 1: Recht, Gesellschaft und Anderes

Die Datenquantität war allgemein niedrig auf Amtsebene. Für den ersten Schritt in der Direktion der Justiz und des Innern gibt es nur 340 Datenpunkte[6], was nicht ausreicht, um ein ML-Modell zu trainieren. Im Vergleich dazu haben wir beispielsweise auf Direktionsebene 8470 valide Datenpunkte. Die Accuracy ist bei diesem Algorithmus dementsprechend schlecht mit 73% und die Resultate unbrauchbar. Ein Upsampling bringt hier keine Verbesserung und reduziert die Accuracy auf 70%. Grund dafür ist die starke Verzerrung bei der Kategorie «Anderes» mit nur 36 Fällen also etwa einem Viertel der anderen beiden Kategorien. Die breite Verzerrung im Zusammenspiel mit der geringen Menge an Daten führt wahrscheinlich dazu, dass das Upsampling nicht hilft.

[5] Im Gespräch wurden fünf Möglichkeiten genannt. Während Recht und Gesellschaft die meisten Geschäfte abdecken, sind die anderen drei Themen (IT, HR und Controlling) eher selten betroffen, was wir auch in der Datenmenge gesehen haben. Daher war eine Vereinfachung für den Algorithmus in nur drei Kategorien sinnvoll.
[6] Diese Zahl muss aufgrund von Data Cleaning mitunter weiter reduziert werden, wodurch einige Datenpunkte wegfallen.

Themen-gruppe	Spezifische Accuracy ohne Upsampling	Anzahl im Datenset ohne Upsampling	Anzahl nach Upsampling	Spezifische Accuracy mit Upsampling
Gesellschaft	73,5%	147	188	69,1%
Recht	60,3%	138	204	57,4%
Anderes	85,3%	36	148	83,8%

Tabelle 2: Resultate auf Amtsebene der Direktion der Justiz und des Innern Schritt 1

Schritt 2: Verteilung an alle zuständigen Ämter und an andere Direktionen
Die Datenquantität für den zweiten Schritt ist mit über 4300 Datenpunkten schon besser, aber die Resultate zeigen, dass hier ebenfalls nicht genug Daten verfügbar sind, insbesondere, da es sich um 27 verschiedene Ämter oder Direktionen handelt.
Hier muss man hinzufügen, dass als Modell alle Ämter und Personen einschliesslich derjenigen Personen aus dem ersten Schritt verwendet wurden. Man könnte die obigen Personen auch aus dem zweiten Schritt ausschliessen. Der Grund aber für den Nichtausschluss ist technischer Natur: Durch eine sigmoide Aktivierung in mehreren möglichen korrekten Ämtern gibt es keinen Verlust für die zusätzliche Vorhersage der drei Kategorien in Tabelle 2. Mit anderen Worten: Es schadet nicht, und man erhält zusätzliche Informationen. Im Prinzip berechnet man damit genau das Gleiche, als wenn man den ersten Schritt nicht hätte, sondern über alle Datenpunkte der Direktion der Justiz und des Innern die Berechnung anstellen würde. Zur Sicherheit haben wir einen Ausschluss ebenfalls gerechnet, aber die Resultate zeigen, wie erwartet, keine nennenswerten Unterschiede gegenüber den Vorhersagen.

Die Resultate zeigen für diese Berechnungen eine deutlich bessere Accuracy als im ersten Schritt: 90,36%. Man muss jedoch auch sagen, dass die Genauigkeit der verschiedenen Ämter zwischen 67% und 98,9% schwankt. Vier von den insgesamt 27 Direktionen und Ämtern haben eine Accuracy von unter 85% bzw. zehn unter einer Accuracy von 90%. Gerade die verschiedenen niedrigeren Werte um 70% und 80% zeigen, dass der Algorithmus nicht sehr zuverlässig ist. Statt 27 Ämter und Direktionsschnittstellen einzeln und mit wenig Mehrwert aufzulisten, zeigt die folgende Tabelle die Ämter mit den schlechtesten und besten Vorhersagen, um einen Eindruck des Spektrums der Resultate zu geben. Die Anzahl von 15 ohne Upsampling zeigt auch, dass man einer Genauigkeit von 98,9% nicht trauen darf, da es ein sehr zufälliges Resultat sein könnte.

Amt	Spezifische Accuracy ohne Upsampling	Anzahl im Datenset ohne Upsampling	Anzahl nach Up-sampling	Spezifische Accuracy mit Upsampling
Amt mit schlechtester Vorhersage	67,0%	278	6805	63,6%
Amt mit bester Vorhersage	98,9%	15	995	98,9%

Tabelle 3: Resultate auf Amtsebene der Direktion der Justiz und des Innern Schritt 2

3.3.2.2 Staatskanzlei

Wie unter 3.2.1 beschrieben, war eine Verknüpfung der alten und neuen GEVER-Daten anhand der Kantonsratsnummer nicht möglich. Somit standen nur die Titel der Geschäfte für ein Training zur Verfügung. Als Proof of Concept konnten wir so trotzdem zeigen, dass eine Triage auch mit sehr wenigen Geschäften funktionieren kann. Wir haben den ersten Ansatz angewendet und konnten ein überraschend gutes Resultat von 93% erreichen.
Der Wert ist allerdings nicht sehr aussagekräftig, da sich die verfügbaren 148 Datenpunkte auf die 8 Abteilungen innerhalb der Staatskanzlei verteilen. Die Anzahl der Geschäfte innerhalb der Abteilungen hat eine Spannbreite von 3 bis 24. Dies zeigt die teilweise geringe Datenquantität. Ein Upsampling bringt hier auch keine Verbesserung. Es reduziert die Gesamtaccuracy auf 91,7% und zeigt eine grundsätzlich schlechtere statt besserer Genauigkeit pro Abteilung. Die Spannbreite nach dem Upsampling betrug 24 bis 30.

Abteilung	Spezifische Accuracy ohne Upsampling	Anzahl im Datenset ohne Upsampling	Anzahl nach Upsampling	Spezifische Accuracy mit Upsampling
Abteilung mit schlechtester Vorhersage	86,7%	21	24	83,3%
Abteilung mit bester Vorhersage	100%	13	24	100%

Tabelle 4: Resultate auf Abteilungsebene der Staatskanzlei

3.3.2.3 Kurze methodische Bemerkung zum Upsampling in den beiden amtsinternen Schritten

Da jeder Datenpunkt aus mehreren korrekten Ämtern oder Abteilungen besteht, kann das Upsampling auch nicht so leicht durchgeführt werden wie auf Direktionsebene. Konkret haben wir die Anzahl derjenigen Ämter oder Abteilungen mit der tiefsten Häufigkeit versucht zu erhöhen. Das benötigt eine Selektion der Datenpunkte, die diese Ämter oder Abteilungen umfassen. Aber diese Datenpunkte umfassen nicht nur diese Ämter oder Abteilungen, sondern auch andere. Deshalb führt ein Upsampling dieser Ämter oder Abteilungen auch zu zwei Effekten: Erstens erhöht dies auch die Ämter oder Abteilungen mit einer grösseren Häufigkeit. Zweitens ist somit die Anzahl der verschiedenen Ämter oder Abteilungen nicht gleichmässig verteilt – teilweise wäre dies sogar logisch nicht möglich. Dadurch kann man die Verzerrung zwischen den verschiedenen Ämtern oder Abteilungen reduzieren, aber nicht aufheben. Beispielsweise reduziert sich die Verzerrung beim ersten oben gezeigten Schritt von 36:147 auf 148:204 – noch dazu tritt nun eine andere Abteilung als häufigere Abteilung auf, was ein ungewollter Effekt eines solchen Upsampling ist.
Methodisch könnte dieser Punkt bei einer Anwendung noch grössere Aufmerksamkeit bekommen, und Vorschläge zur Verbesserung sind da sehr willkommen. Pragmatisch gesehen, könnte es auch Sinn machen, zumindest im zweiten Schritt die Anzahl der möglichen 27 Ämter und Abteilungen zu reduzieren, sodass sich die Verzerrung der Häufigkeiten reduziert.

3.3.3 Ergebnisse für semantische Nähe

3.3.3.1 Versuchsanordnung

Eines der grössten Probleme bei der Verwendung eines «zero-shot learning»-Transformer-Modells ist die Berechnungsdauer. Bei dem zentralen Versuch, alle 129 Themenblöcke der VOG RR einem Inputgeschäft zuzuordnen, beträgt die Berechnungsdauer mehrere Minuten (siehe Resultate in Tabelle 5 etwa vier Minuten im Durchschnitt, aber für einige Geschäfte mit längeren Texten dauerte die Berechnung durchaus sechs Minuten und mehr). Das wäre für eine Anwendung in einer Produktivumgebung definitiv zu lang. Es muss aber unterstrichen werden, dass wir auf unserer IT-Infrastruktur eine sehr moderate CPU-Rechenleistung verwendet haben und der Server keine GPU integriert hatte. Erfahrungsgemäss sollte die Berechnung mit einer GPU mindestens zehnmal schneller sein.

Das Hauptproblem bei der Berechnungsdauer sind die 129 Themenblöcke: Der Algorithmus muss bei einem «zero-shot learning» nämlich 129-mal durchgeführt werden. Weniger Gruppierungen führen zu einer geringeren Berechnungsdauer. Deshalb werden bei einem zweiten, eher experimentellen Versuch, die einzelnen Elemente jeder Direktion in Textblöcke pro Direktion verschmolzen. Das verkürzt die Berechnungsdauer erheblich (auf etwa 18 Sekunden). Der Versuch lässt aber einen interessanten Vergleich mit der Genauigkeit der Vorhersage des ersten Ansatzes basierend auf dem Training der Daten der Kantonsratsgeschäfte zu. Die Resultate in Tabelle 5 zeigen aber, dass die Vorhersage durch eine Aneinanderreihung von Textelementen nicht sehr zielführend ist.

Die Reduktion von 129 in acht Elemente im zweiten Versuch ist sehr kompakt und vermengt semantisch sehr unterschiedliche Themen. Deshalb wurden in einem dritten Versuch die 129 Themenblöcke zu 35 Textgruppierungen zusammengesetzt, die thematisch ähnlich sind. Die Zuordnung war willkürlich und diente dazu, eine Art Zwischenstufe zwischen dem ersten Versuch mit 129 Elementen und dem zweiten Versuch mit acht Elementen zu schaffen.
Die Berechnungsdauer lag bei 74 Sekunden im Schnitt.

3.3.3.2 Performancemessung

Die Versuche zur semantischen Nähe sind, wie bereits erwähnt, stark experimenteller Natur. Das zeigt sich auch in der Performancemessung: Für die Ermittlung der Performance der korrekten semantischen Nähe gab es keine Standardlösung, die wir anwenden konnten. «Zero-shot learning» beruht auf einem vortrainierten Transformers-Modell, um semantische Zusammenhänge zu verstehen. Die Genauigkeit beim Textverständnis hat einen starken Einfluss auf die Genauigkeit der darauf basierten Anwendung. Aber die Genauigkeit dieses Modells hat keinen unmittelbaren Zusammenhang mit der Genauigkeit unseres Anwendungsfalls. Des Weiteren limitieren die Rahmenbedingungen die automatisierte Ermittlung der richtigen Themen, da die Zugehörigkeit des richtigen Themas nirgendwo definiert ist. Wohl können wir aber überprüfen, ob die richtige, d.h. die federführende Direktion ermittelt wurde. Dabei können auch mehrere Direktionen und Ämter einen inhaltlichen Zusammenhang mit dem Geschäft haben.

Als Konsequenz haben wir eine eigene Methode erstellt, um die Accuracy zumindest zu «schätzen». Dabei verwenden wir den Wert der berechneten Konfidenz aus dem Algorithmus und filtern Werte mit einer Wahrscheinlichkeit von über 70%. Je nach Fall können dabei mehrere Themenelemente einen Wert von über 70% haben. Es kann auch sein, dass gar kein Element über dieser Schwelle liegt. Im nächsten Schritt ermitteln wir automatisiert, ob bei denjenigen Themen mit einem Wert von über 70% auch mindestens eines ist, das auch bei der korrekten zuständigen Direktion angesiedelt ist. Denn diese Information ist unsere einzig valable «Ground Truth», mit der wir die Resultate automatisiert vergleichen können. Das ist mit mehreren interpretativen Schwierigkeiten verbunden, weshalb wir nur von einer «Schätzung» sprechen können. Beispielsweise werden auch Fälle als korrekt eingestuft, bei denen ein Thema der

richtigen Direktion mit einer hohen Wahrscheinlich zugewiesen wird, das aber in Wirklichkeit inhaltlich nichts mit dem Thema des Geschäfts zu tun hat. Das wäre falsch-positiv. Als Konsequenz muss man davon ausgehen, dass die wahre Genauigkeit niedriger ist.

Ebenfalls könnte man einwenden, dass man als methodische Alternative nur dasjenige Thema mit der höchsten Wahrscheinlichkeit verwenden könnte. Diese methodische Entscheidung ist induktiv begründet: Das Filtern von jeweils zwei bis drei Themen scheint uns in den verschiedenen Experimenten eine plausible Methode zu sein, ohne dass das Thema mit dem höchsten Wert immer das aussagekräftigste Thema zum dazugehörenden Geschäft war.
Ebenfalls ist der Wert von 70% sehr willkürlich und als Resultat von Versuchen mit mehreren Schwellenwerten gewählt.

Die folgende Tabelle zeigt die Resultate bezüglich der geschätzten Accuracy, die wir mit den genannten drei Methoden ermitteln konnten. Die geschätzte Accuracy liegt bei 73%, wenn alle 129 Themenblöcke der VOG RR verwendet werden. Es dauert aber erstaunliche vier Minuten im Schnitt, um eine Berechnung durchzuführen.

Versuch	Geschätzte Accuracy	Durchschnittlich benötigte Zeit für Berechnung
alle 129 Themenblöcke der VOG RR	73%	241 Sekunden
35 thematische Blöcke	28%	74 Sekunden
Textblöcke pro 8 Direktionen	28%	18 Sekunden

Tabelle 5: Resultate der geschätzten Accuracy

Bei den anderen beiden Varianten mit weniger Textblöcken wurde eine gleich niedrige geschätzte Accuracy von 28% erreicht. Die geringe Accuracy überrascht nicht, da der semantische Zusammenhang dabei aus einer Aneinanderreihung von inhaltlich sehr unterschiedlichen Elementen ermittelt wird. Überraschend ist jedoch, dass die geschätzte Accuracy in beiden Varianten übereinstimmt.
Wie zu erwarten hat die Berechnungsdauer einen direkten Zusammenhang mit der Anzahl der verwendeten Blöcke. Die Experimente zeigen aber, dass es sich nicht lohnt, die Genauigkeit der Informationen in den einzelnen Themenblöcken künstlich zu reduzieren.

3.3.3.3 Weitere Ideen

Es gibt noch weitere Ideen, die aber mangels Zeit nicht getestet werden konnten. Erfolgversprechend könnte «few-shot learning» statt «zero-shot learning» sein.[7]

Eine Möglichkeit wäre beispielsweise, einen Zwischenschritt mit einer Textzusammenfassung zwischen dem Inputtext und dem Zero-Shot-Algorithmus einzufügen.[8] Momentan scheinen aber die automatisierten Textzusammenfassungen noch keine sehr gute Leistung zu haben.

[7] siehe z.B. towardsdatascience.com/sentence-transformer-fine-tuning-setfit-outperforms-gpt-3-on-few-shot-text-classification-while-d9a3788f0b4e
[8] Siehe z.B. towardsdatascience.com/overcoming-input-length-constraints-of-transformers-b0dd5c557f7e

4 Erkenntnisse aus dem Pilotprojekt

Das Pilotprojekt zeigt anhand eines konkreten Anwendungsbeispiels, wie Machine Learning in der Verwaltung einen Mehrwert schaffen kann. Entscheidungsunterstützende Algorithmen können Prozesse, hier spezifisch die Triage, optimieren. Das Projekt beweist experimentell die Machbarkeit der Grundidee. Um den Piloten weiterentwickeln und produktiv einsetzen zu können, wurde ein Handlungsbedarf in verschiedenen Themenbereichen[9] erkannt. Dieser lässt sich auf den Einsatz anderer Machine-Learning-Ansätze in der kantonalen Verwaltung übertragen und wird in Kapitel 5 mit Massnahmen konkretisiert.

4.1 Organisation

Die involvierten Fachpersonen der Verwaltung haben positiv auf das Projekt reagiert. Im Verlauf des Projekts haben sie zahlreiche weitere mögliche Anwendungsfälle für entscheidungsunterstützende Algorithmen in ihrem Arbeitsalltag aufgebracht, die mit meist repetitiven wie auch teilweise unnötigen Teilprozessen zusammenhängen.

Diese Reaktionen zeigen eindrücklich, dass im Arbeitsalltag die Integration von Prozess- und Datenwissen fehlt. Um Verwaltungsabläufe mithilfe der Digitalisierung verbessern, vereinfachen oder automatisieren zu können, brauchen Teams neben ihrem Fachwissen ebenfalls Daten- und Statistikkompetenz sowie ein Verständnis für die zugrunde liegenden Datenprozesse. Eine enge interdisziplinäre Zusammenarbeit zwischen Fachspezialistinnen und -spezialisten und Datenspezialistinnen und -spezialisten stellt sicher, dass ein Verbesserungspotenzial erkannt und die beste Lösung gefunden wird. Die Zusammenarbeit fördert darüber hinaus die Datenkompetenz der Fachpersonen und das Verständnis der vollständigen Datenprozesse der Datenspezialistinnen und -spezialisten während ihrer täglichen Arbeit.

4.2 Daten

Machine Learning ist datenintensiv. Um dessen Potenzial zu erkennen und den Einsatz nachzuvollziehen, ist eine grundlegende Daten- und Statistikkompetenz nötig. Bereits das Pilotprojekt hat gezeigt, dass ein grösseres Verständnis von möglichen Anwendungen Künstlicher Intelligenz Mitarbeitende befähigt, in ihrem Arbeitsalltag das Potenzial für Verbesserungen zu benennen. Datenkompetenz führt so zu mehr Digitalisierungsvorschlägen «von unten». Ein Vorgehen, das dank der Expertise der Fachpersonen Akzeptanz und Effizienz verspricht.

Die Realisierung des Pilotprojekts hat die hohen Ansprüche an Datenqualität von Machine-Learning-Anwendungen verdeutlicht. Merkmale einer hohen Datenqualität sind Einheitlichkeit, Zuverlässigkeit, Eindeutigkeit, Vollständigkeit, Aktualität und Korrektheit.[10] Obwohl in der kantonalen Verwaltung viele Daten erhoben und bearbeitet werden, sind sie oft weder vergleichbar noch über Amts- und Direktionsgrenzen hinweg verknüpfbar, weil einheitliche Vorgaben (Standards und Definitionen) noch fehlen, was einige Datenanwendungen schlicht verunmöglicht. Eine konsistente Dateneingabe und -bearbeitung verbessert die Qualität der für entscheidungsunterstützende Algorithmen zur Verfügung stehenden Daten und ermöglicht somit den Einsatz von aussagekräftigen und verlässlichen Systemen.

Datenanwendungen wie in diesem Pilotprojekt stärken beides: sie erhöhen die Daten- und Statistikkompetenz und zeigen auf, dass die Vorteile einer besseren Datenqualität den Nachteil des grösseren Initialaufwands deutlich überwiegen.

---

[9] Die gewählten Themenbereiche sind denjenigen im RRB Nr. 1362/2021 angelehnt.
[10] Mehr zu Datenqualität: zh.ch/de/politik-staat/statistik-daten/datenkompetenz/data-und-statistical-literacy.html#-968727060

4.3 Infrastruktur

Damit Machine-Learning-Anwendungen nicht nur experimentell bleiben, sondern produktiv genutzt werden können, muss die Infrastruktur einige Rahmenbedingungen erfüllen: Daten müssen in hoher Qualität aus den bestehenden Systemen exportiert und deren Anwendung in bestehende Systeme integriert werden können. Dazu sind standardisierte, programmierbare Schnittstellen (API, Application Programming Interfaces) notwendig. Zusätzlich braucht es eine Infrastruktur, die den benötigten Rechen- und Serverleistungen gerecht wird. Das bedingt zahlreiche langfristige Erwägungen im Zusammenhang mit der IT-Infrastruktur (siehe Kapitel 5.3).

Im konkreten Triage-Beispiel stand die GEVER im Mittelpunkt. Die Mitarbeitenden verschiedener Direktionen arbeiten täglich damit und führen zahlreiche Arbeitsschritte direkt in dieser Umgebung aus, was für eine Prozessoptimierung zwingend berücksichtigt werden muss. Für das Pilotprojekt haben wir Daten für das Training eines entscheidungsunterstützenden Algorithmus genutzt und hätten dann diese Anwendung wieder in die Arbeitsumgebung des GEVER integrieren wollen, damit die Mitarbeitenden diese Triage-Vorschläge direkt in ihrer alltäglichen Arbeitsumgebung eingebettet erhalten. Diese Innovation bedingt somit, dass solche Arbeitsumgebungen offen und adaptiv sind. Da die GEVER-Lösung proprietär ist und nicht auf offenem Quellcode basiert, führt dies zu «Lock-Ins»[11]. Eine Implementierung benötigt zwangsweise das Zutun der Herstellerfirma. Rasche, inkrementelle wie auch experimentelle Lösungsfindungen sind erschwert.

Zudem ist keine geeignete IT-Infrastruktur für eine Produktionsumgebung vorhanden. Wir konnten für dieses Pilotprojekt mit dankenswerter Unterstützung der IT-Abteilung der Direktion der Justiz und des Innern einen experimentellen Server mit minimalen technischen Voraussetzungen aufsetzen, um eine interaktive Lösung zu gestalten. Diese Infrastruktur ermöglicht jedoch kein fundierteres Trainieren von ML-Algorithmen, das u.a. Hardwarekomponenten wie GPU benötigen würde.

4.4 Recht und Ethik

Eine nutzendenzentrierte, effiziente und effektive Verwaltung muss ihre Prozesse laufend verbessern. Wenn die Prozesse mit Machine Learning verbessert werden sollen, stellen sich rechtliche und ethische Fragen, wie sie bereits in der KI-Studie[12] diskutiert wurden. Der im Bericht vorgeschlagene Gesetzesrahmen für Anwendungen Künstlicher Intelligenz in oder durch die Verwaltung bildet die benötigten Leitplanken für diese Verbesserungen. In allen Projekten, die den Einsatz von KI zur Folge haben, muss darüber hinaus der Rechtsgrundlagenanalyse nach Hermes hohes Gewicht beigemessen werden. Wo die Analyse ergibt, dass rechtliche Anpassungen notwendig sind, sollten diese Anpassungen technologieneutral formuliert und die Ziele und Konsequenzen in den Mittelpunkt gestellt werden. Ausserdem ist zu prüfen, ob für künftige Vorhaben im Bereich der Künstlichen Intelligenz übergreifende neue rechtliche Grundlagen geschaffen werden sollen, damit die Analyse in Einzelfällen einfacher wird.

[11] Siehe z.B. Brown, Fishenden und Thompson (2014) «Digitizing Governments – Understanding and Implementing New Digital Business Model».
[12] zh.ch/de/news-uebersicht/medienmitteilungen/2021/04/kuenstliche-intelligenz-in-der-verwaltung-braucht-klare-leitlinien.html

Das Pilotprojekt wurde bewusst so gewählt, dass weder sein Datenbedarf noch seine Auswirkungen rechtliche oder ethische Risiken nach sich ziehen würden. Die in der KI-Studie vorgestellten Checklisten führten im Pilotprojekt dementsprechend nicht zu einer höheren Risikoerwägung, weil drei wichtige Faktoren bei der Auswahl es Pilotprojekts beachtet wurden:

- Die für das Training vorgesehenen Daten enthalten keine schützenswerten Merkmale.
- Die Anwendung zielt auf eine Entscheidungsunterstützung, und nicht auf eine vollautomatisierte Triage.
- Eine falsche Triageentscheidung führt nicht zu schwerwiegenden Konsequenzen, sondern meist zu einer Korrektur seitens der attribuierten Direktion.

Dennoch hat die Anwendung der in der KI-Studie vorgestellten Checklisten zu einer Sensibilisierung des Teams geführt. Die frühe Konfrontation mit rechtlichen und ethischen Fragen hat so eine positive Auswirkung auf die Nachhaltigkeit des Projektergebnisses.

Die Checklisten ermöglichen eine Risikoabwägung und ermöglichen durch die Transparenz, die ethischen Prinzipien zu erreichen. Auch experimentelle Projekte wie dieses Pilotprojekt profitieren von einem Eintrag in einem öffentlichen KI-Register, wie es vom Kantonsrat in einem Postulat gefordert wurde und vom Regierungsrat nun umgesetzt wird[13] und wie es auf Bundesebene bereits umgesetzt wurde[14]. Die frühe und transparente Publikation von KI-Projekten in der Pilotphase stärkt die Zusammenarbeit mit Forschungseinrichtungen und zivilgesellschaftlichen Akteuren. Das verbessert nicht nur die technischen Lösungen, sondern führt auch dazu, dass frühzeitig ethische Fragestellungen aufgeworfen und geklärt werden.

Um die rechtlichen und ethischen Abklärungen vor der Umsetzung von KI-Projekten weiter zu stärken, benötigt die Verwaltung eine direktionsübergreifende Anlaufstelle. Neben der Beratung bei rechtlichen Grundlagen und den ethischen Implikationen der Anwendung muss diese Anlaufstelle auch die Daten auf verschiedenartige Benachteiligungen überprüfen und algorithmischen Aspekte beurteilen. Dazu ist die enge Zusammenarbeit zwischen Juristinnen und Juristen mit Datenwissenschaftlerinnen und -wissenschaftlern und der Einbezug von Ethikexpertinnen und -experten notwendig. Das jeweilige Fachwissen muss eng miteinbezogen werden.

5 Nächste Schritte und nötige Massnahmen

Die technische Machbarkeit wie auch die Rückmeldungen von Mitarbeitenden lassen ein positives Fazit aus dem Pilotprojekt für die Anwendung von entscheidungsunterstützenden Algorithmen zu. Das Pilotprojekt konnte auch die Herausforderungen darlegen, die bewältigt werden müssen, um KI-Projekte in produktive Anwendungen einer digitalen Verwaltung der Zukunft überzuführen.

Die weiteren Schritte im Pilotprojekt sind überschaubar: Um die Anwendung in die bestehende Oberfläche des GEVER der Direktion der Justiz und des Innern und der Staatskanzlei integrieren zu können, müssen grundlegenden Punkte geklärt werden:

- Das Einverständnis der innerhalb der Staatskanzlei und der Direktion der Justiz und des Innern zuständigen Personen für die Weiterentwicklung der GEVER-Systeme muss eingeholt werden.
- Basierend auf einer Offerte müssen die Verantwortlichen für die jeweiligen GEVER-Systeme in der Staatskanzlei und der Direktion der Justiz und des Innern im Anschluss an eine Kosten-Nutzen-Analyse entscheiden, ob das Pilotprojekt produktiv umgesetzt werden soll.
- Dies umfasst die Klärung der Finanzierung der Weiterentwicklung.

[13] Siehe Postulat KR-Nr. 9/2022 betreffend Transparenz über den Einsatz von künstlicher Intelligenz in der Verwaltung, kantonsrat.zh.ch/geschaefte/geschaeft/?id=7eaa64973b454a6ca4e5eceb86b13941
[14] Auf Bundesebene wurde das KI-Register durch das Kompetenznetzwerk für künstliche Intelligenz veröffentlicht, siehe cnai.swiss

Die Ausbreitung des Piloten auf die gesamte Verwaltung birgt die Herausforderung, dass in den Direktionen und Ämtern eine heterogene GEVER-Landschaft besteht. Die Systeme sind oft proprietär und werden von externen Anbietern weiterentwickelt. Für eine kantonsweite Ausbreitung des Piloten ist eine Standardisierung dieser Systeme unabdingbar, wie sie im Vorhaben IP6.6 (Koordinierte Ausbreitung der verwaltungsinternen elektronischen Geschäftsabwicklung) aus dem Impulsprogramm vorgesehen ist und im Rahmen eines geplanten Nachfolgeprojekts ab 2023 weitergeführt werden soll. Aus Sicht des KI-Piloten ist anzustreben, dass die Systeme quelloffen sind und Weiterentwicklungen von Schnittstellen von internen wie auch externen Stellen durchgeführt werden können.

Über das konkrete Pilotprojekt hinaus zeigt sich ein Handlungsbedarf für eine konstruktive, effiziente und gewinnbringende Anwendung von KI-Lösungen in der Verwaltung auf genereller Ebene, wie in Kapitel 4 beschrieben. Um diese Herausforderungen meistern zu können, empfiehlt das Projektteam aufgrund der Erfahrungen mit der Pilotanwendung Massnahmen in verschiedenen Themenbereichen.

5.1 Organisation

- Verstärkte Zusammenarbeit zwischen Fach- und Datenspezialistinnen und -spezialisten in interdisziplinären Projektteams. Somit kann die «Zusammenarbeit und das ganzheitliche Denken bei der Leistungserbringung» verbessert werden, wie in den Leitsätzen «gemeinsam digital unterwegs»[15] gefordert.
- Machine Learning und Ansätze der künstlichen Intelligenz können einen Einfluss darauf haben, wie wir arbeiten und wie wir zusammenarbeiten. Diese Aspekte sollen in der Entwicklung einer Organisation berücksichtigt werden. Hierfür müssen neue Kompetenzen gefördert und Mitarbeitende geschult werden. Im Rahmen der strategischen Initiative Organisation gibt es Handlungsfelder und Ansätze, die solche Entwicklungen aufgreifen können.

5.2 Daten

- Stärkung der Daten- und Statistikkompetenz aller Verwaltungsmitarbeitenden mit dem Ziel, den Nutzen von Daten, Machine Learning und gemeinsamer Anwendungen zu verstehen und deren Umsetzung aktiv unterstützen zu können. Diese Massnahme ist auch in den Leitsätzen «gemeinsam digital unterwegs» verankert.[16]
- Daten sind eine strategische Ressource.[17] Ein Wissens- und Praxisaustausch entlang von Datenprozessen, Fachthemen und IT-Applikationen («communities of practice») führt zu einer stärkeren Koordinierung zwischen unterschiedlichen Rollen und zwischen unterschiedlichen Verwaltungseinheiten. Dies wiederum steigert das Potenzial für Auswertungen, Anwendungen und Verwendbarkeit (z.B. Verknüpfungen von verschiedenen Datensätzen). Die mit der Umsetzung der strategischen Initiative Daten betrauten Verwaltungseinheiten bündeln das notwendige datenwissenschaftliche Knowhow und stellen es den anderen Verwaltungseinheiten zur Verfügung.
- Schaffen eines KI-Registers und frühzeitige, transparente Publikation von Projekten bereits in der Pilotphase. Das Ziel ist, dank Rückmeldungen aus der Forschung und der Zivilgesellschaft sowohl bessere technische Umsetzungen als auch ethisch unbedenkliche Ergebnisse zu erhalten.

[15] Siehe S.5 Themenbereich Organisation, RRB Nr. 1362/2021.
[16] «Die Datenkompetenz in der Verwaltung wird nachhaltig gefördert und gemeinsame Standards erleichtern die Nutzung der Daten.» siehe S.6 unter Themenbereich Daten, RRB Nr. 1362/2021.
[17] Siehe auch Leitsätze «gemeinsam digital unterwegs», S.6 Themenbereich Daten, RRB Nr. 1362/2021.

5.3 Infrastruktur

- Beschaffung von IT-Infrastruktur mit primärem Fokus auf offenen Quellcode, um Lock-Ins zu verhindern und die Innovation zu fördern. Falls dies nicht möglich ist, können offene Schnittstellen eine Weiterentwicklung ermöglichen.
- Adaptive Systeme, um Anpassungen der Arbeitsumgebung einfach und verwaltungsintern vornehmen zu können.
- Flexible und stabile (Server-)Infrastruktur, damit experimentelle Projekte auch in eine Produktionsumgebung übergeführt werden können.
- Eine ausreichende Infrastruktur für das Trainieren von Machine Learning (z.B. mit Hardwarekomponenten wie GPU). Diese Anforderungen haben sich auch im Rahmen der Innovation-Sandbox ergeben, die von der Standortförderung des Kantons initiiert wurde.[18]

5.4 Recht und Ethik

- Rechtsgrundlagenanalyse bei KI-Projekten grosses Gewicht einräumen und daraus erfolgende Gesetzesanpassungen technologieneutral formulieren.
- Schaffen einer direktionsübergreifenden Anlaufstelle zu rechtlicher und ethischer Beurteilung von KI-Projekten. Diese Stelle sorgt dafür, dass diese Projekte aus rechtlicher, ethischer, statistischer und fachlicher Sicht beurteilt werden.
- Gegebenenfalls Veröffentlichung der Algorithmen selbst, falls keine Rückschlüsse auf die Einzeldaten oder Geschäftsgeheimnisse möglich sind.
- Veröffentlichung eines Transparenzberichtes auch für risikoarme Anwendungen.

5.5 Leistungen

- Werden Leistungen der kantonalen Verwaltung im Rahmen der digitalen Transformation neu entwickelt, geschieht dies unter Berücksichtigung neuer technischer Möglichkeiten. Hier sollen automatisierte Ansätze systematisch mitgeprüft werden – idealerweise als festes Element im Framework der Leistungsentwicklung. Dies betrifft sowohl Leistungen an Bevölkerung und Wirtschaft als auch verwaltungsinterne Leistungen.

5.6 Schlussfolgerung

Als digitale Verwaltung der Zukunft wollen wir uns künftig aktiv mit Themen der Künstlichen Intelligenz auseinandersetzen, weitere Anwendungsfälle suchen, weitere Pilote durchführen und diese zu einer Good Practice wachsen lassen. Die Strategieumsetzung im Rahmen der strategischen Initiativen sowie das politische Interesse am Thema legitimieren die Auseinandersetzung mit dem Thema und helfen, dieses in der Verwaltung mehr und mehr zu verankern.

[18] zh.ch/de/wirtschaft-arbeit/wirtschaftsstandort/innovation-sandbox.html

6 Anhang

6.1 Abbildungsverzeichnis

6.2 Tabellenverzeichnis