# Einleitung

Wie können Unternehmen ihre Beschaffungsprozesse effizienter und kostengünstiger gestalten, während sie gleichzeitig die Einhaltung komplexer Beschaffungsrichtlinien sicherstellen? Diese Frage wird angesichts der steigenden Anzahl von Produkten und Dienstleistungen, die täglich von großen Unternehmen beschafft werden müssen, immer drängender. Die manuelle Bearbeitung von Freitextanfragen, das Verständnis komplexer Anforderungen und die Lenkung der Ausgaben entsprechend bestehenden Verträgen und Katalogen sind zeitaufwendig und fehleranfällig. In diesem Kontext bietet unsere Lösung auf Basis eines Siamese Netzwerks eine innovative Möglichkeit, diese Herausforderungen zu meistern. Durch die Anpassung des Embedding Spaces für Produktbeschreibungen mittels eines vortrainierten SimCSE Modells können wir die automatische Verarbeitung von Beschaffungsanfragen verbessern und somit die Effizienz und Genauigkeit der Beschaffungsprozesse erheblich steigern.

Obwohl unsere Lösung nur ein Teil der gesamten Automatisierungslösung ist, spielt sie eine entscheidende Rolle dabei, Unternehmen dabei zu unterstützen, ihre Beschaffungsprozesse zu optimieren und die Einhaltung von Beschaffungsrichtlinien sicherzustellen.

# Theoretischer Hintergrund

## Vorhersage von Satzähnlichkeit

Die Ähnlichkeit zweier Sätze mathematisch zu messen, gilt seit ca. 1957 als Ziel im Forschungsbereich von Natural Language Processing (NLP). Dazu wurden verschiedenste Ansätze verfolgt, doch in den meisten Fällen waren die Ergebnisse aufgrund des Mangels geeigneter Trainingsdaten unzureichend. Dies liegt daran, dass in kaum einer alltäglichen oder geschäftlichen Tätigkeit die Ähnlichkeit zwischen zwei Sätzen im großen Stil erfasst wird. Erst mit dem Aufkommen von großen Sprachmodellen bzw. deren Word Embeddings konnte der Durchbruch erzielt werden.

### Word Embeddings

Moderne Sprachmodelle werden mit Milliarden Trainingsdaten trainiert und sind dadurch in der Lage die natürliche Sprache zu verstehen und zu generieren. Zum Training wird den Modellen ein Satz mit einem fehlenden Wort gegeben, das das Modell dann vorhersagen muss. Dadurch lernt das Modell den semantischen Inhalt von Wörtern und deren inhaltliche Verbindung zu anderen Wörtern. Neben der Funktion zur Vorhersage von fehlenden Wörtern eines Textes kann das Modell den semantischen Inhalt eines Wortes in einen mathematischen Vektor überführen. Diesen mathematischen Vektor nennt man Word Embedding.

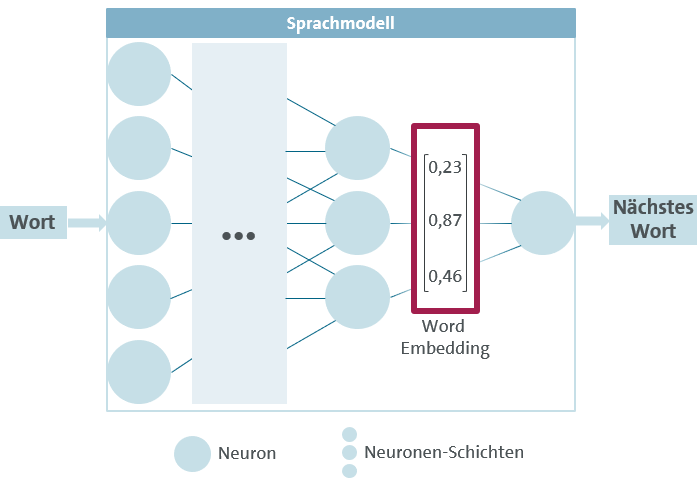


Abbildung 1: Prozess zur Erstellung eines Word Embeddings  
Quelle: Eigene Abbildung

Die Erstellung eines Word Embeddings wird in Abbildung 1 dargestellt. Zunächst wird das entsprechende Wort in das Sprachmodell eingegeben und über mehrere Schichten des Modells verarbeitet. Das Sprachmodell komprimiert von Schicht zu Schicht den Informationsgehalt des Wortes. Dabei hat jeder Bereich des Modells durch das Training gelernt, auf spezifische Informationen zu achten und diese zu repräsentieren. In der letzten Schicht sind die Informationen in der komprimiertesten Form und werden genutzt, um das nächste Wort vorherzusagen. Die Werte der letzten Schicht können jedoch auch in einen Vektor überführt werden, der alle semantischen Informationen des Wortes in höchst komprimierter Form repräsentiert.

### Sentence Embeddings

Für einige Anwendungsfälle im NLP-Bereich sollen jedoch nicht nur Wörter, sondern ganze Sätze in einen mathematischen Vektor umgewandelt werden. Zur Erstellung dieser Vektorrepräsentationen werden die einzelnen Wörter eines Satzes in das Modell eingegeben, das jeweilige Word Embedding extrahiert und mittels des Durchschnittes zu einem gemeinsamen Vektor bzw. einem sogenannten Sentence Embedding zusammengeführt. Die Repräsentationen von ausschließlich vortrainierten Modellen sind jedoch für die Praxis nicht anwendbar. Dies hängt damit zusammen, dass das Modell nicht dazu ausgelegt ist den Inhalt der Wörter bzw. Sätze im Vektorraum zu verteilen, sondern dieser nur ein Nebenprodukt der Wortvorhersage ist. Zur besseren Verteilung und Abgrenzung von Sätzen im Vektorraum muss das Modell mithilfe der Siamese Netzwerk Architektur nachtrainiert werden.

### Siamese Netzwerk Architektur

In der Siamese Netzwerk Architektur werden die Wörter von zwei Sätzen in das vortrainierte Sprachmodell eingegeben und die jeweiligen Word Embeddings extrahiert. Mittels des Durchschnittes werden die Word Embeddings in ein Sentence Embedding zusammengeführt. Anschließend wird die Cosinus Distanz der zwei Vektoren bestimmt (siehe Abbildung 2). Zum Trainieren des Netzwerkes erhalten sehr ähnliche Satzpaare den Wert eins und verschiedene Satzpaare den Wert null. Nachdem das Modell für das jeweilige Satzpaar die Sentence Embeddings bzw. die Cosinus Distanz berechnet hat, wird das Modell entsprechend der Abweichung vom erwarteten und tatsächlichen Ergebnis. Dies sorgt dafür, dass ähnliche Sätze im Vektorraum sehr nah beieinander liegen, während unterschiedliche Sätze weit voneinander entfernt sind.

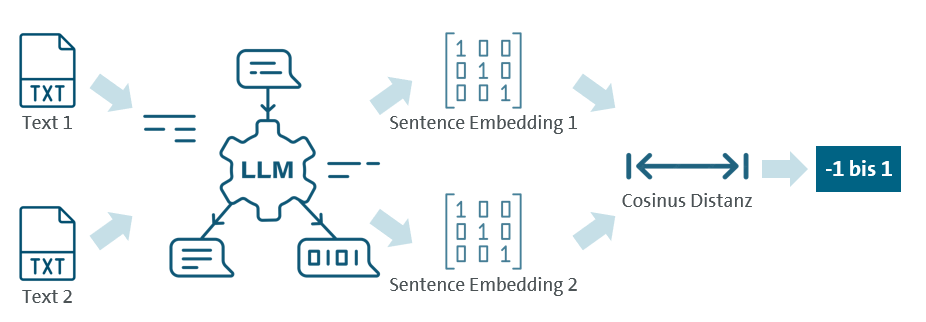


Abbildung 2: Siamese Netzwerk-Architektur  
Quelle: Eigene Abbildung

Das Nachtrainieren eines vortrainierten Sprachmodells führt zu exzellenten Ergebnissen und ist aktuell das erfolgreichste Verfahren zur Vorhersage von Satzähnlichkeiten. Die größte Herausforderung des Ansatzes ist jedoch die Sammlung und Aufbereitung der entsprechenden Trainingsdaten. Es müssen immer Satzpaare gebildet werden, die vom Inhalt sehr ähnlich oder sehr unterschiedlich sind. Dieser Aufbau von Trainingsdaten ist aufwendig, aber der entscheidende Erfolgsfaktor für gute Sentence Embeddings.

# Modellentwicklung

Um unser Modell zu Trainieren verwenden wir einen Kaggle Datensatz, welcher bestimmte Informationen über Amazon Produkte enthält. Hierunter fällt der Name des Produktes, eine Beschreibung, Stichpunkte, die die wichtigsten Informationen über das Produkt beschreiben und eine Produkt Typ ID. Nach dem Bereinigen des Datensatzes wurden positive Trainingspaare und negative Trainingspaare erstellt. Die Positiven Trainingspaare werden erstellt indem zwei Produkte aus der gleichen Kategorie gewählt werden mit den Voraussetzungen, dass maximal zehn positive Paare pro Kategorie gewählt werden und, dass innerhalb der Kategorien zufällige Stichproben ausgewählt werden, um eine gewisse Diversität zu garantieren. Die negativen Trainingspaare werden erstellt, indem zwei Produkte aus verschiedenen Kategorien ausgewählt werden. Die Anzahl an positiven und negativen Paaren ist gleich, damit kein Bias während dem Training des Models vorhanden ist. Die Trainingsdaten sind aufgeteilt in ca. 10% Testdaten und ca. 90% Trainingsdaten

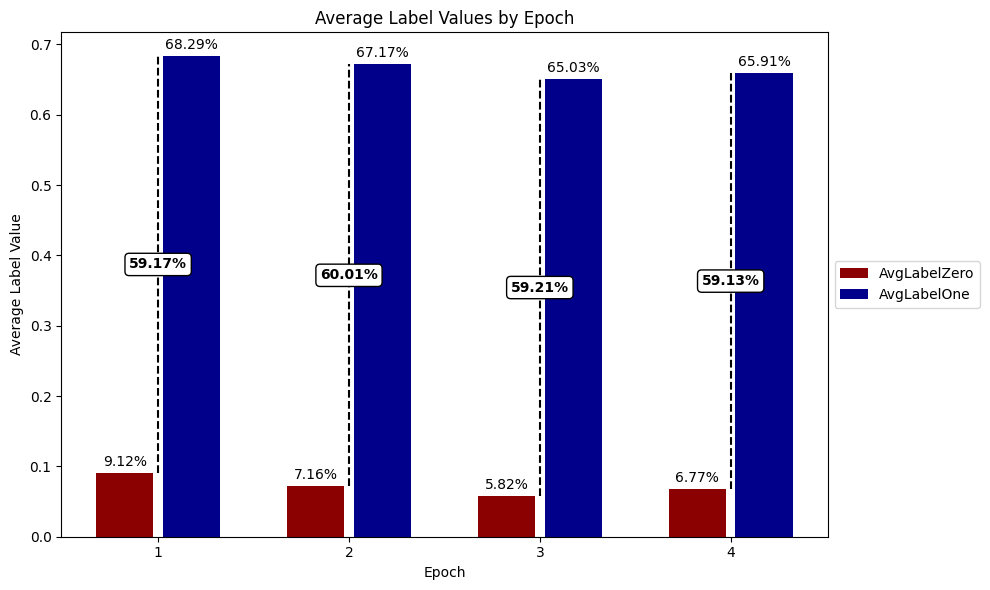


Abbildung 3: Trainingsergebnisse je Epoche  
Quelle: Eigene Abbildung

Verwendet wurde das SimCSE Model von Princeton basierend auf der RoBERTa-Large, welches spezifisch für Satzähnlichkeit vortrainiert ist. Um das Ergebnis des Models darzustellen, wird die Cosinus Distanz zwischen den beiden Produktbeschreibungen ausgerechnet. Es werden separat die Cosinus Distanzen für die positiven als auch die negativen Paare ausgerechnet und im Anschluss visualisiert. Nachdem das Model 4 Epochen trainiert wurde, war das Ergebnis, dass in Epoche zwei bereits der beste Wert von 60.01% erreicht war und der Wert in den anderen drei Epochen leicht über 59% befand. Nachdem die Ergebnisse tiefer analysiert wurden, stellte sich jedoch heraus, dass die Kategorisierung von Amazon teilweise nicht besonders Aussagekräftig ist und führt zu der Erkenntnis, dass eine detaillierte Kategorisierung für die Produkte ein essenzieller Faktor für die Performance des Models ist.

# Business Case

## Marktperspektive und Relevanz

Der Markt für Einkaufsautomatisierung ist derzeit enorm und wächst rasant. Zahlreiche Startups und etablierte Unternehmen beschäftigen sich intensiv mit der Entwicklung innovativer Lösungen zur Optimierung von Beschaffungsprozessen. Diese Lösungen basieren häufig auf fortschrittlichen Technologien wie maschinellem Lernen, Big Data und Blockchain. Ziel ist es, die Bereinigung und Klassifizierung von Beschaffungsdaten zu automatisieren und Unternehmen eine umfassende und konsolidierte Sicht auf ihre Beschaffungsausgaben zu bieten.

## Konkretes Problem

Mitarbeiter in großen Unternehmen haben oft Schwierigkeiten, die komplexen Beschaffungsregeln zu verstehen und einzuhalten, wenn sie Produkte oder Dienstleistungen erwerben möchten. Dies führt zu tausenden von Anfragen an die Beschaffungsteams, die oft als Freitext eingereicht werden. Diese Freitextanfragen sind mit den aktuellen Methoden kaum automatisch zu verarbeiten, was bedeutet, dass jede einzelne Anfrage manuell bearbeitet werden muss.

Die manuelle Bearbeitung dieser Anfragen ist äußerst zeitaufwendig und kostspielig. Beschaffungsteams verbringen unzählige Stunden damit, die Anforderungen der Mitarbeiter zu verstehen, die Ausgaben entsprechend bestehenden Verträgen und Katalogen zu lenken, falsche oder fehlende Informationen zu korrigieren und eine Ausgabenanalyse durchzuführen. Dies macht den gesamten Beschaffungsprozess ineffizient und teuer.

## Breite der möglichen Anwendungsfälle

Unsere Lösung mit der Siamese Netzwerk Architektur bietet vielfältige Anwendungsmöglichkeiten, die weit über die reine Automatisierung der Beschaffungsanfragen hinausgehen. Zum Beispiel:

* **Vergleichbare Angebote erkennen:** Unternehmen können unsere Lösung nutzen, um ähnliche Produkte und Dienstleistungen auf verschiedenen Plattformen zu identifizieren und zu vergleichen. Dies ermöglicht eine bessere Markttransparenz und hilft Einkäufern, die besten Angebote zu finden und Kosten zu sparen.
* **Web Scraping und Datenintegration:** Durch die Integration mit einem Webscraper können Beschaffungsteams kontinuierlich aktuelle Produktinformationen und Preise von verschiedenen Anbietern sammeln. Unser Modell kann dann diese Daten verarbeiten und vergleichbare Angebote automatisch identifizieren. Diese Kombination ermöglicht es Unternehmen, eine stets aktuelle und umfassende Marktübersicht zu haben.
* **Vergleichsplattformen:** Plattformen wie Idealo setzen ähnliche Technologien ein, um Nutzern zu helfen, die besten Angebote zu finden. Unsere Lösung könnte in solchen Plattformen eingesetzt werden, um die Genauigkeit und Effizienz der Vergleichsalgorithmen zu verbessern und somit den Nutzern bessere und schnellere Suchergebnisse zu bieten.
* **Vertrags- und Katalogmanagement:** Unsere Technologie kann dazu beitragen, bestehende Verträge und Kataloge zu verwalten und zu optimieren, indem sie ähnliche Produkte innerhalb dieser Dokumente identifiziert. Dies unterstützt die Einhaltung von Verträgen und erleichtert die Pflege und Aktualisierung von Katalogen.

## Beispiel eines Startups

Ein Beispiel für ein Startup, das sich mit diesem Problem beschäftigt, ist [Asklio](https://de.asklio.ai/). Dieses Unternehmen entwickelt eine generative AI-basierte Softwarelösung, die es Beschaffungsteams ermöglicht, Freitextanfragen effizient zu verarbeiten. In der idealen Anwendung können Mitarbeiter ihre Anfragen in natürlicher Sprache stellen und sofort sehen, in welchem Katalog sie die benötigten Produkte finden oder bei welchen Anbietern bereits bestehende Verträge bestehen.

Unsere Lösung mit der Siamese Netzwerk Architektur passt perfekt in dieses Szenario. Sie ermöglicht es, Produkte anhand ihrer Beschreibungen zu unterscheiden und somit gleiche Produkte auf unterschiedlichen Webseiten zu identifizieren. Dies unterstützt die Automatisierung der Verarbeitung von Freitextanfragen und verbessert die Effizienz und Genauigkeit der Beschaffungsprozesse erheblich.

Durch die Implementierung unserer Lösung können Unternehmen ihre Beschaffungsprozesse drastisch beschleunigen und gleichzeitig die Einhaltung von Beschaffungsrichtlinien verbessern. In einem dynamischen Marktumfeld, in dem zahlreiche Startups nach effizienten Lösungen suchen, bieten wir eine robuste und skalierbare Technologie, die den modernen Anforderungen der Beschaffungsteams gerecht wird.

# Fazit

Unser Projekt hat eine äußerst effektive Lösung im Bereich der natürlichen Sprachverarbeitung hervorgebracht, die eine breite Palette von Anwendungsfällen abdeckt. Durch die Entwicklung und Anpassung eines Siamese Netzwerks zur Verbesserung des Embedding Spaces für Produktbeschreibungen konnten wir eine Methode implementieren, die präzise und effiziente Vergleiche zwischen Produktbeschreibungen ermöglicht.

Unsere Lösung zeigt, dass fortschrittliche NLP-Techniken wie das vortrainierte SimCSE Modell auf RoBERTa-Basis erheblich zur Verbesserung der Satzähnlichkeit beitragen können. Die Anwendung dieser Techniken bietet vielfältige Einsatzmöglichkeiten, von der Automatisierung von Textverarbeitungsaufgaben bis hin zur Optimierung von Informationssystemen.

Das Projekt demonstriert eindrucksvoll, wie moderne Machine Learning-Methoden im Bereich der natürlichen Sprachverarbeitung eingesetzt werden können, um komplexe Herausforderungen zu bewältigen. Es bildet eine solide Grundlage für weiterführende Forschungsarbeiten und praktische Anwendungen, die über den ursprünglichen Anwendungsfall hinausgehen und in verschiedenen Branchen und Bereichen zur Verbesserung der Effizienz und Genauigkeit beitragen können.