

- Lorenz, J. H. (2003). *Lernschwache Rechner fördern. Ursachen der Rechenschwäche. Frühhinweise auf Rechenschwäche. Diagnostisches Vorgehen*. Cornelsen.
- Lorenz, J. H. & Radatz, H. (1993). *Handbuch des Förderens im Mathematikunterricht*. Schroedel.
- Pielsticker, F. (2020). *Mathematische Wissensentwicklungsprozesse von Schülerinnen und Schülern. Fallstudien zu empirisch-orientiertem Mathematikunterricht mit 3D-Druck*. Springer Spektrum.
- Pielsticker, F. & Witzke, I. (2022). Erkenntnisse zur Beschreibung des aktivierten mathematischen Wissens in empirischen Kontexten an einem Beispiel aus der Wahrscheinlichkeitstheorie. *Mathematica Didactica*, 45. DOI:10.18716/ojs/md/2022.1395
- Pollak, H. & Garfunkel, S. (2014). A View of mathematical modeling in mathematics education. In: A. Sanfratello & B. Dickmann (eds.), *Proceedings of Conference on Mathematical Modeling at Teachers College of Columbia University* (S. 6–12). New York.
- Rolka, K., & Halverscheid, S. (2011). Researching young students' mathematical world views. *ZDM Mathematics Education*, 43, 521–533. DOI:10.1007/s11858-011-0330-9
- Schipper, W. (2009). *Handbuch für den Mathematikunterricht an Grundschulen*. Schroedel.
- Schoenfeld, A. (1985). *Mathematical Problem Solving*. Academic Press.
- Söbbeke, E. & Steinbring, H. (2004). Was ist Mathematik? – Vorstellungen von Grundschulkindern. In P. Scherer, & D. Bönig (Hrsg.), *Mathematik für Kinder – Mathematik von Kindern* (Bd. 117, S. 26–38). Grundschulverband – Arbeitskreis Grundschule e. V.
- Stake, R. E. (1995). *The Art of Case Study Research*. Sage Publications.
- Vollrath, H. J. (1989). Funktionales Denken. *Journal für Mathematik-Didaktik*, 10, 3–37. DOI:10.1007/BF03338719
- Wartha, S. & Schulz, A. (2019). *Rechenproblemen vorbeugen* (6. Aufl.). Cornelsen.
- Yin, R. K. (2018). *Case Study Research and Applications. Design and Methods* (6<sup>th</sup> ed.). Sage Publications.

Jenny Knöppel, Universität Siegen  
knoepfel@mathematik.uni-siegen.de

Dr. Felicitas Pielsticker, Universität Siegen  
pielsticker@mathematik.uni-siegen.de

## Mathematik und Sprache – Textanalyse im Mathematikunterricht

Mohini Nonnenmann, Martin Vogt, Simone Bast und Karsten Lübke

### Einleitung

„Eines der mächtigsten Instrumente der Menschheitsgeschichte“ ([6]), „Jede Lehrkraft muss ChatGPT kennen“ ([2]): Das sind nur zwei der zahlreichen Titel aus bekannten Onlinezeitungen zum Thema ChatGPT, einem Chatbot der menschenähnliche Texte verfassen kann. Die Titel unterstreichen eindrucksvoll die Bedeutung der Erstellung und Auswertung von Texten, denn wir kommen täglich mehrfach damit in Kontakt, etwa bei der Zeitungslektüre, in Büchern, in Social Media oder auf Webseiten im Internet.

Für zahlreiche Fragestellungen und Anwendungen ist es hilfreich, Texte automatisch auswerten zu können, zum Beispiel um verwendete Sprache von Politikerinnen und Politikern, in Büchern oder in Zeitungen miteinander vergleichen zu können.

Andere Beispiele betreffen Unternehmen, etwa bei der Beantwortung der Frage: Was wird über mein Unternehmen und meine Produkte im Internet geschrieben?

Deshalb sollten Texte und deren Auswertung auch im Mathematikunterricht behandelt werden. Dabei können Texte für verschiedene Themen des Lehrplans als Datengrundlage verwendet werden. Schließlich basiert auch ChatGPT auf der statistischen Analyse riesiger Textmengen.

Im Folgenden möchten wir die Möglichkeiten der Textanalyse im Mathematikunterricht anhand der Lernsituation eines Praktikums in der Marketingabteilung des fiktiven Sportartikelherstellers *TurboKick*<sup>1</sup> aufzeigen:

*Lernsituation: Sie absolvieren ein Praktikum beim Sportartikelhersteller TurboKick und haben die Aufgabe bekom-*

<sup>1</sup> Der Name wurde mithilfe von ChatGPT erstellt.

men, geeignete Influencer für eine Marketingkampagne zu identifizieren. Hierzu sollen Sie eine Liste geeigneter Influencer erstellen, die sich mit dem Thema Sport beschäftigen und deren verwendete Sprache zu TurboKick passt. Damit die Analyse in Zukunft wiederholt werden kann, soll die Auswahl automatisch anhand der verwendeten Themen und Wörter in bereits veröffentlichten Artikeln der Influencer erfolgen.

Um die Aufgabe zu bearbeiten, werden wir zunächst kurz skizzieren, wie Texte zur inhaltlichen Analyse aufbereitet werden können. Anschließend geben wir drei Beispiele für die Verwendung von Texten, im Rahmen von funktionalen Zusammenhängen, der Wahrscheinlichkeitsrechnung, sowie bei der Vektoren- und Matrizenrechnung.

Bei der Auswahl der Beispiele war es uns wichtig, eine möglichst große Bandbreite an Themen aus dem Lehrplan abzudecken.<sup>2</sup> Bei der Lerngruppe handelt es sich um eine Klasse aus der Oberstufe eines Beruflichen Gymnasiums in Rheinland-Pfalz.

### Textaufbereitung zur weiteren Analyse

Texte sind sogenannte unstrukturierte Daten. Das bedeutet, dass diese, im Gegensatz zu Daten in Tabellen, keinem vorgefertigten Datenmodell folgen und vor der Analyse zusätzlich aufbereitet werden müssen. Typische Beispiele hierfür sind, eine Korrektur der Rechtschreibung, die Entfernung von Satzzeichen oder die Vereinheitlichung von Groß- und Kleinschreibung. Weitere Möglichkeiten sind die Rückführung von Wörtern auf ihre Wortstämme (z. B. Familien zu Familie) und die Entfernung von häufig vorkommenden Wörtern, wie *der*, *die*, *und*, etc., welche wenig zum Inhalt beitragen (sogenannte Stoppwörter). Für den Mathematikunterricht müssen nicht alle Maßnahmen durchgeführt werden. Die folgenden Beispiele können auch ohne diese Vorbereitungen verwendet werden. Sollen aber größere Texte unter Verwendung des Computers ausgewertet werden, etwa im Rahmen eines fächerübergreifenden Unterrichts mit der Informatik, empfiehlt es sich die oben erwähnten Vorbereitungen anzusprechen.<sup>3</sup> Im Folgenden gehen wir davon aus, dass die Texte bereits geeignet aufbereitet worden sind.

### Funktionale Zusammenhänge mit Wörtern

*Lernaufgabe 1:* Sie haben die Aufgabe bekommen, automatisch geeignete Influencer für eine Marketingkampagne zu identifizieren. Die Auswahl soll automatisch

anhand der verwendeten Themen und Wörter erfolgen. Dabei stellt sich Ihnen die Frage: Sind alle Wörter zur Unterscheidung von Texten gleich wichtig?

Zahlen und Funktionen helfen, die Realität zu beschreiben. Dies gilt auch für Texte. Zentraler Bestandteil der Analyse von Texten ist es herauszufinden, welche Wörter ein Dokument besonders gut beschreiben, die also für Text und Inhalt charakteristisch sind.

Die folgenden Sätze beinhalten beide das Wort *feiert*, aber der Inhalt ist unterschiedlich:

Dokument 1:

„WM-Geschäft – Adidas feiert WM-Sieg“ ([4])

Dokument 2:

„Wie Europa den Wahlsieg Macrons feiert.“ ([7])

In diesem Beispiel ist etwa das Wort *Wahlsieg* besonders wichtig, um es der Dokumentenklasse Politik zuzuordnen und von Dokument 1 zu unterscheiden, während das Wort *feiert* nicht hilfreich ist, denn es kommt in beiden Texten gleichhäufig vor. Zur Charakterisierung von Texten sind also Wörter, welche häufig in wenigen Dokumenten auftreten, wichtiger als Wörter welche eine hohe Frequenz in der gesamten Textsammlung aufweisen. Das statistische Maß TF-IDF (Term Frequency-Inverse Document frequency) formalisiert diese Idee:

$$\text{TF-IDF}_{i,j} = \text{TF}_{i,j} \cdot \text{IDF}_i \quad (1)$$

wobei  $\text{TF}_{i,j}$  die absolute Häufigkeit von Wort  $i$  in Dokument  $j$  misst und  $\text{IDF}_i$  die inverse Dokumenthäufigkeit für Term  $i$  darstellt, welche berechnet wird als ([10, S. 52]):

$$\text{IDF}_i = \ln \left( \frac{n}{\text{DF}_i} \right)$$

Hier ist  $n$  die Gesamtzahl an Dokumenten, während  $\text{DF}_i$  angibt, in wie vielen Dokumenten der Sammlung, der Term  $i$  vorkommt.

Für das Wort *WM* in Dokument 1 ergibt sich demnach der Wert 1,38:

$$\text{TF-IDF}_{\text{WM},\text{D}_1} = 2 \cdot \ln \left( \frac{2}{1} \right) = 2 \cdot 0,69 = 1,38$$

Hingegen erhält das Wort *feiert* den Wert 0:

$$\text{TF-IDF}_{\text{feiert},\text{D}_1} = 2 \cdot \ln \left( \frac{2}{2} \right) = 2 \cdot 0 = 0$$

Insgesamt ergibt sich die Tabelle 1.

Aus der Tabelle wird ersichtlich, dass jede Zeile ein Dokument repräsentiert, somit jedes Dokument

<sup>2</sup> [https://berufsbildendeschule.bildung-rp.de/fileadmin/user\\_upload/bbs/berufsbildendeschule.bildung-rp.de/Lehrplaene/Dokumente/Lehrplan\\_2014/2015-01-08\\_LP\\_BG\\_Mathe.pdf](https://berufsbildendeschule.bildung-rp.de/fileadmin/user_upload/bbs/berufsbildendeschule.bildung-rp.de/Lehrplaene/Dokumente/Lehrplan_2014/2015-01-08_LP_BG_Mathe.pdf)

<sup>3</sup> Das Buch [www.tidytextmining.com](http://www.tidytextmining.com) bildet eine gute Basis für Textanalyse mit dem Programm R.

Tabelle 1. TF-IDF Matrix

	WM	Geschäft	Sieg	Adidas	feiert	Europa	Macron	Wahlsieg
D1	1,38	0,69	0,69	0,69	0,00	0,00	0,00	0,00
D2	0,00	0,00	0,00	0,00	0,00	0,69	0,69	0,69

als Vektor betrachtet werden kann. Die Spalten bilden die TF-IDF gewichteten Wörter ab. In Zeile 1 befindet sich ein Zeitungsartikel aus der Rubrik Sport, wohingegen das Dokument in Zeile 2 aus der Rubrik Politik stammt. Besonders hoch wurde das Wort WM gewichtet, denn es kommt in diesem Dokument 2 mal vor, während es in Dokument 2 nicht erscheint. Des weiteren fällt auf, dass das Wort *feiert* einen TF-IDF Wert von null hat, denn es kommt in allen Dokumenten gleich häufig vor und besitzt somit keine Relevanz für die Unterscheidung.

Die inverse Dokumenthäufigkeit gewichtet die Ausdrücke in Abhängigkeit ihrer Frequenz. Durch die Multiplikation mit der Term-Frequenz, ergibt sich ein individueller TF-IDF Wert für jedes Wort (vgl. Formel 1). Der Logarithmus  $\ln$  dient hier als Glättungsfunktion, da bei sehr niedrigen  $DF_i$  Werten kleine Unterschiede große Auswirkungen auf  $\frac{n}{DF_i}$  haben können ([9]).

Stellen wir die inverse Dokumenthäufigkeit als Funktion dar, so ist  $DF_i$  die unabhängige Variable und der IDF Wert die abhängige Variable. Die Anzahl an Dokumenten wird durch  $n$  angegeben, welche als Konstante betrachtet werden kann. In Abbildung 1 wird angenommen, dass die Textsammlung aus 100 Dokumenten besteht. Zu sehen ist der Funktionsgraph von  $(\frac{n}{DF_i})$  und  $\ln(\frac{n}{DF_i})$ . Es fällt auf, dass der Graph von  $\ln(\frac{n}{DF_i})$  für kleine Werte von  $DF_i$  deutlich glatter verläuft. Somit erhalten Wörter, die in wenigen Dokumenten vorkommen, ein vergleichsweise geringeres Gewicht.

Im Lehrplan für das Fach Mathematik an berufsbildenden Schulen<sup>4</sup> kann diese erste Lernaufgabe im Lernbereich 1: „Denken in funktionalen Zusammenhängen“ verortet werden. Die in diesem Lernbereich ausgewiesenen Kompetenzen erlangen die Lernenden, indem sie in die Rolle der Praktikantin bzw. des Praktikanten schlüpfen und funktionale Zusammenhänge im Hinblick auf die Worthäufigkeit und die inverse Dokumenthäufigkeit darstellen, analysieren und interpretieren. Konkret würde sich hier eine Herangehensweise frei nach dem Grundprinzip des kooperativen Lernens ([3]) (Think – Pair – Share) anbieten. In der Think-

Phase könnten sich die Lernenden in Einzelarbeit anhand unterschiedlicher kurzer Texte das Konzept der Worthäufigkeit erarbeiten und diese in Form einer Tabelle (vgl. Tabelle 1) darstellen. Während der sich anschließenden Pair-Phase könnten die Ergebnisse aus der Think-Phase zusammengeführt (vgl. Tabelle 1), diskutiert, grafisch dargestellt (vgl. Abb. 1) und anhand dieser Darstellung verglichen werden. Die abschließende Share-Phase dient der Präsentation der Ergebnisse der Kleingruppen. In Bezug zur Lernsituation könnten die Lernenden im Anschluss einen Leitfaden für die Analyse von Texten erstellen, den sie ihren Kolleginnen und Kollegen im Rahmen des Praktikums zur Verfügung stellen, damit auch diese in der Lage sind, Texte im Hinblick auf die Worthäufigkeit zu analysieren. Zur weiteren Vertiefung der mathematischen Kompetenzen könnten dann zum Beispiel die Vor- und Nachteile einer logarithmischen Darstellung von Häufigkeiten diskutiert werden.

Auch ermöglicht die vektorielle Darstellung von Dokumenten oder Wörtern weiterführende Anwendungen (siehe auch Tabelle 1 und Abschnitt 8), denn es können Distanzmaße auf die Vektoren angewendet werden. Mittels dieser Distanzmaße kann dann die Ähnlichkeit zwischen Dokumenten berechnet werden. Sind Dokumente im Vektorraum weit voneinander entfernt, weisen sie eine geringe Ähnlichkeit zueinander auf, sind sie nah beieinander weisen sie eine hohe Ähnlichkeit auf ([1, S. 77]).

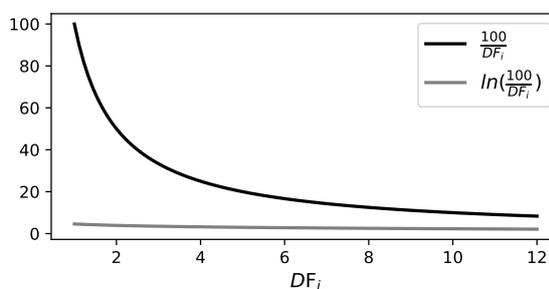


Abbildung 1. IDF und Logarithmus

<sup>4</sup> [https://berufsbildendeschule.bildung-rp.de/fileadmin/user\\_upload/bbs/berufsbildendeschule.bildung-rp.de/Lehrplaene/Dokumente/Lehrplan\\_2014/2015-01-08\\_LP\\_BG\\_Mathe.pdf](https://berufsbildendeschule.bildung-rp.de/fileadmin/user_upload/bbs/berufsbildendeschule.bildung-rp.de/Lehrplaene/Dokumente/Lehrplan_2014/2015-01-08_LP_BG_Mathe.pdf)

## Wahrscheinlichkeitsrechnung mit Wörtern

*Lernaufgabe 2: Nachdem Sie sich mit der Bedeutung einzelner Wörter in Texten auseinandergesetzt haben, möchten Sie als nächstes geeignete Influencer identifizieren. Dazu schauen Sie sich bereits verfasste Texte von Influencern an. Wie können Texte automatisch der Kategorie Sport zugeordnet werden?*

Für zahlreiche Anwendungen ist es wichtig, Texte automatisch anhand ihres Inhalts einer Klasse bzw. eines Themas zuzuordnen, also zu klassifizieren, etwa eine E-Mail in Spam oder nicht-Spam oder einen Zeitungsartikel in die Kategorie Politik oder Sport.

Im Folgenden soll das Dokument mit den Worten: „Argentinien feiert WM-Titel“ ([12]), der Klasse Politik oder der Klasse Sport unter Zuhilfenahme der Wahrscheinlichkeitsrechnung, zugeordnet werden. Ein gängiges Verfahren hierfür ist die so genannte Naive Bayes-Klassifikation, welche auf dem Satz von Bayes beruht ([5, S. 249]):

$$P(\text{Klasse}|\text{Daten}) = \frac{P(\text{Daten}|\text{Klasse}) \cdot P(\text{Klasse})}{P(\text{Daten})} \quad (2)$$

In diesem Beispiel nehmen wir an, dass uns bereits insgesamt 10 000 Dokumente aus den Bereichen Politik und Sport vorliegen, also andere Zeitungsartikel, auf deren Grundlage die Klassifikation erfolgt. Davon gehören 4000 Dokumente zur Kategorie Politik und 6000 zur Kategorie Sport. Weiter nehmen wir an, dass 300 Artikel der Klasse Sport das Wort *Argentinien* enthalten, 200 das Wort *feiert*, 600 das Wort *WM* und 350 das Wort *Titel*. Es gilt demnach:

$$P(\text{Politik}) = \frac{4.000}{10.000} = 0,4$$

$$P(\text{Sport}) = \frac{6.000}{10.000} = 0,6$$

Zudem lässt sich die bedingte Wahrscheinlichkeit der einzelnen Wörter unter der Bedingung, dass die Klasse schon bekannt ist folgendermaßen ermitteln:

$$P(\text{Argentinien}|\text{Sport}) = \frac{300}{6.000} = 0,05$$

$$P(\text{feiert}|\text{Sport}) = \frac{200}{6.000} = 0,033$$

$$P(\text{WM}|\text{Sport}) = \frac{600}{6.000} = 0,1$$

$$P(\text{Titel}|\text{Sport}) = \frac{350}{6.000} = 0,058$$

Aus den bedingten Wahrscheinlichkeiten der einzelnen Wörter, ermittelt sich die Wahrscheinlichkeit

für ein Dokument, welches genau diese Wörter enthält und dessen Klassenzugehörigkeit bereits bekannt ist, gemäß der folgenden Formel:

$$P(\text{Dokument}|\text{Klasse}) = P(\text{Wort}_1|\text{Klasse}) \cdot P(\text{Wort}_2|\text{Klasse}) \cdot \dots \cdot P(\text{Wort}_i|\text{Klasse})$$

Hier wird angenommen, dass alle Merkmale unabhängig voneinander auftreten. Aus diesem Grund wird das Verfahren als *naiv(e)* bezeichnet, denn diese Annahme ist in der Realität meistens verletzt ([5], S. 251). Beispielhaft hierfür können die Wörter *WM* und *Finale* herangezogen werden, welche oft in Zusammenhang miteinander auftreten. Dennoch zeigt der Naive Bayes-Klassifikator in der Praxis zufriedenstellende Ergebnisse und wird häufig verwendet ([5], S. 251).

In unserem konkreten Fall bedeutet dies:

$$\begin{aligned} P(\text{Dokument}|\text{Sport}) &= \\ P(\text{Argentinien}|\text{Sport}) \cdot P(\text{feiert}|\text{Sport}) \cdot \\ P(\text{WM}|\text{Sport}) \cdot P(\text{Titel}|\text{Sport}) &= \\ 0,05 \cdot 0,033 \cdot 0,1 \cdot 0,058 &= 0,00000957 \end{aligned}$$

Nun lässt sich die Wahrscheinlichkeit für die Klassenzugehörigkeit Sport, nach dem Satz von Bayes, wie folgt berechnen:

$$\begin{aligned} P(\text{Sport}|\text{Dokument}) &= P(\text{Dokument}|\text{Sport}) \cdot \\ P(\text{Sport}) &= 0,00000957 \cdot 0,6 = 0,000005742 \end{aligned}$$

Hier wurde nicht wie in der oben angegebenen Formel (vgl. Formel 2) durch  $P(\text{Daten})$  geteilt, da dieser Berechnungsschritt die Klassifikationsentscheidung nicht beeinflusst und nur als Skalierungsfaktor dient.

Endgültig wird ein Dokument mithilfe der folgenden Entscheidungsregel klassifiziert ([8]):

*Wenn  $P(\text{Sport}|\text{Dokument}) \geq P(\text{Politik}|\text{Dokument})$  klassifiziere das Dokument in die Kategorie Sport, andernfalls in die Kategorie Politik.*

Für das vorliegende Beispiel wird das Dokument in die Kategorie Sport klassifiziert:<sup>5</sup>

$$\begin{aligned} P(\text{Sport}|\text{Dokument}) &= 0,000005742 \\ &\geq P(\text{Politik}|\text{Dokument}) = 0,00000031 \quad (3) \end{aligned}$$

Anhand dieses Beispiels wurde dargelegt, wie eine Klassifikation mit Wörtern als Eingabewerte unter Zuhilfenahme der Wahrscheinlichkeitsrechnung funktioniert.

Die zweite Lernaufgabe kann im Lehrplan für das Fach Mathematik an berufsbildenden Schulen

<sup>5</sup> Unter der Annahme, dass 100 Artikel der Kategorie Politik das Wort *Argentinien*, 200 das Wort *feiert*, 100 das Wort *WM* und 100 das Wort *Titel* enthalten.

im Lernbereich 6: „Untersuchen von Zufällen und Analysieren von Daten“ verortet werden. Hierbei erlangen die Lernenden die vorgeschlagenen Kompetenzen, indem sie Beiträge unterschiedlicher Influencer vergleichen und bewerten, inwieweit diese Influencer zum Sportartikelhersteller bzw. zum Thema Sport passen. Hierbei stellt dieser Vergleich die Anwendung des zuvor erlernten Konzepts der bedingten Wahrscheinlichkeit dar. Auch hier könnten anhand des kooperativen Lernens zunächst in Einzelarbeit (Think) verschiedene Influencer analysiert, in Kleingruppen verglichen (Pair) und schließlich mehrere passende Influencer ausgewählt werden, nachdem man in der Präsentationsphase einen Überblick über alle Influencer gewonnen hat.

### Vektorrechnung mit Wörtern

*Lernaufgabe 3: Sie haben erfolgreich eine Liste von Sportinfluencern erstellt. Nun möchten Sie diejenigen auswählen, welche auch sprachlich passen, also eine ähnliche Sprache wie TurboKick verwenden. Dabei stellt sich Ihnen die Frage: Wann sind Wörter ähnlich zueinander?*

In Abschnitt 8 wurde bereits aufgezeigt, dass Dokumente als Vektoren dargestellt werden können. Dieses Prinzip macht sich auch das sogenannte Word Embedding zu nutze, indem Wörter als Vektoren dargestellt werden. Im Folgenden soll Word Embedding anhand des Beispielsatzes: „Argentinien Sie bereite große Freude“ erläutert werden.

Wörter besitzen Eigenschaften, wie Sentiment, Häufigkeit und Wortart. Die Wörter Sieg und Freude sind zum Beispiel meistens positiv konnotiert, während Ländernamen neutral belegt sind. Im Word Embedding wird dies wie folgt verwendet: Die Dimension Sentiment sei beispielsweise in negativ = -1, neutral = 0 und positiv = 1 unterteilt. Dadurch erhalten die Wörter Sieg und Freude jeweils eine 1 für diese Dimension, während die anderen als neutral wahrgenommenen Wörter eine 0 erhalten. Eine zweite Dimension könnte die Häufigkeit eines Wortes in Texten sein. Häufigkeiten lassen sich in selten = -1, mittel = 0 und häufig = 1 abstufen. Im Allgemeinen wird das Wort groß/große häufig verwendet und erhält eine 1, während alle anderen Wörter eher mit einer mittleren Häufigkeit vorkommen (z. B. 0). Schließlich könnten Wörter gemäß ihrer Wortart eingeteilt werden. Wortarten gliedern sich unter anderem in Nomen = -1, Adjektive = 0 und Verben = 1. Insgesamt ergibt sich die Kodierung in Tabelle 2.

Tabelle 2. Word Embedding

	Sentiment	Häufigkeit	Wortart
Argentinien	0	0	-1
Sieg	1	0	-1
bereite	0	0	1
große	0	1	0
Freude	1	0	-1

Tabelle 3. Ähnlichkeitsmatrix

	Argen- tiniens	Sieg	bereite	große	Freude
Argentinien	1	1	-1	0	1
Sieg	1	2	-1	0	2
bereite	-1	-1	1	0	-1
große	0	0	0	1	0
Freude	1	2	-1	0	2

Diese Kodierung führt dazu, dass ähnliche Wörter auch ähnliche Vektoren haben und somit nahe beieinander liegen.<sup>6</sup> Zudem wird dadurch die Bedeutung der einzelnen Wörter erfasst. Die Ähnlichkeit zwischen zwei Vektoren lässt sich anhand der Matrixmultiplikation der Matrix aus Tabelle 2 mit der transponierten Version ermitteln ([11]). Das Ergebnis ist die Ähnlichkeitsmatrix aus Tabelle 3.

Das Wort Sieg, mit dem Vektor (1, 0, -1) und das Wort Freude mit dem gleichen Vektor (1, 0, -1) kommen in der Matrix aus Tabelle 3 auf den gemeinsamen Wert 2 (Eintrag 2 Zeile, letzte Spalte) durch Multiplikation der Vektoren:

$$(1, 0, -1) \cdot \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} = 2$$

Schließlich können die Vektoren im Raum dargestellt werden, um Ähnlichkeiten und Abstände zwischen den Wörtern zu berechnen (siehe Abbildung 2).

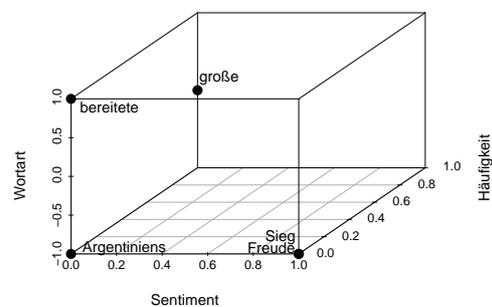


Abbildung 2. 3D-Abbildung der Wortvektoren

<sup>6</sup> Bei dieser einfachen Kodierung erhalten die Wörter Sieg und Freude sogar die gleichen Werte.

In der Praxis werden den Wörtern mehr Eigenschaften bzw. Dimensionen zugeordnet. Zudem geschieht dies nicht manuell, sondern kann auch auf Grundlage vorhandener Dokumente bestimmt werden.

Im Lehrplan für das Fach Mathematik an berufsbildenden Schulen lässt sich die dritte Lernaufgabe dem Lernbereich 3: „Algebraisierung von mehrdimensionalen Verflechtungen und analytische Beschreibung des Raumes“ zuordnen. Zu guter Letzt könnten hier erneut kooperativ Ähnlichkeiten der Sprache der Influencer zur Sprache des Sportartikelherstellers TurboKick identifiziert werden, indem die Lernenden zunächst in Einzelarbeit das Word Embedding (vgl. Tabelle 2) für jeweils einen Influencer erstellen (Think), anschließend in Kleingruppen die Ähnlichkeitsmatrix (vgl. Tabelle 3) erstellen (Pair) und abschließend im Plenum mit den Ähnlichkeitsmatrizen der anderen Kleingruppen vergleichen, um den passenden Influencer auszuwählen.

### Zusammenfassung und Ausblick

*Stolz präsentieren Sie die ausgewählten Influencer Ihrem Team. Gleichzeitig schlagen Sie vor, Texte auch in Zukunft automatisch zu analysieren, etwa um Kritik von Kunden an den Produkten oder neue Trends in der Sportwelt frühzeitig zu erkennen. Ihr Vorschlag wird begeistert aufgenommen.*

Die Bedeutung von Texten als Datentyp nimmt immer weiter zu. Ein Grund hierfür ist die schier unendlich große Menge an verfügbaren Texten in Zeitungen, Büchern, Social Media oder im Internet. Chatbots wie ChatGPT nutzen solche Texte um eine künstliche Intelligenz zu erschaffen. Um dieser Entwicklung gerecht zu werden, ist es wichtig, Texte auch im Mathematikunterricht zu behandeln und in verschiedene Bereiche des Lehrplans einzubinden. Hierzu zählen funktionale Zusammenhänge, Wahrscheinlichkeitsrechnung und Vektoren- bzw. Matrizenrechnung. In diesem Artikel wurden hierzu drei Beispiele anhand der Lernsituation eines Sportartikelherstellers skizziert, die als Grundlage zur Integration von Texten in den Mathematikunterricht dienen können.

Die vorgestellten Beispiele können auf verschiedene Weisen erweitert werden. Ein Beispiel hierfür ist die Berechnung von Wortzusammenhängen bzw. Wortkorrelationen (siehe [10, Kapitel 4]). Hierbei könnten Schülerinnen und Schüler etwa analysieren, welche Wörter Politiker oder Influencer häufig gemeinsam in einem Tweet verwenden. Texte können zudem als Grundlage für fächerübergreifenden Unterricht dienen. Hierfür gibt es zahlreiche Möglichkeiten der fächerübergreifenden Verzahnung,

etwa mit dem Ethikunterricht im Themenblock Medienkompetenz (z. B. Fake News), dem Deutschunterricht durch den automatischen Vergleich von Sprache zwischen verschiedenen Textgattungen oder zwischen Autoren, mit dem Fach Politik durch die oben erwähnte Analyse der Sprache von Politikerinnen und Politikern, dem Fach Geschichte durch die Analyse des Wandels der Sprache im Zeitverlauf oder dem Fach Informatik durch die Analyse von großen Texten mit geeigneter Software wie R oder Python.

### Literatur

- [1] Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. *Practical Text Analytics: Maximizing the Value of Text Data*. Springer International Publishing, Basel, Schweiz, 1 edition, 2018.
- [2] Lisa Becker. Künstliche Intelligenz: Jede Lehrkraft muss ChatGPT kennen, 2023. [tinyurl.com/2p9b8425](https://tinyurl.com/2p9b8425); abgerufen: 20. 4. 2023.
- [3] Ludger Brüning and Tobias Saum. *Erfolgreich unterrichten durch Kooperatives Lernen*. Neue Deutsche Schule Verlagsgesellschaft mbH, Essen, 5 edition, 2009.
- [4] News Deutschland. WM-Geschäft: Adidas feiert WM-Sieg, 12. 2022. [tinyurl.com/wv89vxd9](https://tinyurl.com/wv89vxd9); abgerufen: 28. 4. 2023.
- [5] Detlev Frick, Andreas Gadatsch, Jens Kaufmann, Birgit Lankes, Christoph Quix, Andreas Schmidt, and Uwe Schmitz, editors. *Data Science: Konzepte, Erfahrungen, Fallstudien und Praxis*. Springer Fachmedien Wiesbaden, Wiesbaden, 2021.
- [6] Sascha Lobo. Künstliche Intelligenz Eines der mächtigsten Instrumente der Menschheitsgeschichte. *Spiegel*, 3 2023. [tinyurl.com/yc439d63](https://tinyurl.com/yc439d63); abgerufen: 20. 4. 2023.
- [7] Florian Neuhann. Wie Europa den Wahlsieg Macrons feiert, 4 2022. [tinyurl.com/3yj7fj84](https://tinyurl.com/3yj7fj84); abgerufen: 28. 4. 2023.
- [8] Sebastian Raschka. Naive Bayes and text classification I – Introduction and theory. *CoRR*, abs/1410.5329, 2014. arXiv:1410.5329
- [9] Rudi Seitz. Understanding TF-IDF and BM-25, March 2020. [kmwllc.com/index.php/2020/03/20/understanding-tfidf-and-bm-25/](https://kmwllc.com/index.php/2020/03/20/understanding-tfidf-and-bm-25/); abgerufen: 20. 4. 2023.
- [10] Julia Silge and David Robinson. *Text Mining with R: A Tidy Approach*. O'Reilly Media, 2017.
- [11] Holger von Joanne-Diedrich. Attention! What lies at the Core of ChatGPT? (Also as a Video!) – Learning Machines, 3 2023. [blog.ephorie.de/attention-what-lies-at-the-core-of-chatgpt/](https://blog.ephorie.de/attention-what-lies-at-the-core-of-chatgpt/); abgerufen: 5. 5. 2023.
- [12] Deutsche Welle. Argentinien feiert WM-Titel | DW | 19. 12. 2022, 2022. [tinyurl.com/yxnw3y48](https://tinyurl.com/yxnw3y48); abgerufen: 05.05.2023.

Mohini Nonnenmann, Hochschule Trier  
E-Mail: [mhnn4252@hochschule-trier.de](mailto:mhnn4252@hochschule-trier.de)

Martin Vogt, Hochschule Trier  
E-Mail: [m.vogt@wir.hochschule-trier.de](mailto:m.vogt@wir.hochschule-trier.de)

Simone Bast, Staatliches Studienseminar für das Lehramt an berufsbildenden Schulen Trier  
E-Mail: [simone.bast@bbs-tr.semrlp.de](mailto:simone.bast@bbs-tr.semrlp.de)

Karsten Lübke, FOM Hochschule, Dortmund  
E-Mail: [karsten.luebke@fom.de](mailto:karsten.luebke@fom.de)