Digital Business University of Applied Sciences

Data Science & Management (M.Sc.)

ADS-01: Tools der Softwareentwicklung und Online-Daten

Prof. Dr. Marcel Hebing

Themenmodellierung mittels Latent Dirichlet Allocation auf Zeitungshomepages

Studienarbeit

Zusammenfassung
Die Studie untersucht, ob sich die redaktionellen Strategien von FAZ, SZ und Welt in den veröffentlichten Artikeln widerspiegeln. Während die FAZ vor allem über Politik, Wirtschaft und Kultur berichtet, deckt die SZ ein breites Themenspektrum mit Schwerpunkten in Gesellschaft, Wissenschaft und Feuilleton ab. Die Welt wiederum konzentriert sich stark auf internationale Politik, Wirtschaft und Finanzen. Über einen Zeitraum von drei Monaten wurden täglich die Startseiten der drei Zeitungen mittels Web-Scraping erfasst, so dass insgesamt 276 HTML-Seiten mit 45.143 Artikeln analysiert wurden. Nach der Datenbereinigung wurde ein LDA-Modell trainiert, um thematische Schwerpunkte zu identifizieren. Die Ergebnisse zeigen, dass die FAZ vor allem politische und wirtschaftliche Themen behandelt, während die SZ eine größere Vielfalt aufweist und insbesondere soziale und kulturelle Inhalte betont. Die Welt zeichnet sich durch einen starken Fokus auf internationale Politik, Wirtschaft und Technologie aus. Die Hypothesen, dass die FAZ vor allem über Politik und Wirtschaft berichtet, die SZ eine größere Themenvielfalt mit den Schwerpunkten Gesellschaft und Wissenschaft aufweist und die Welt stärker über internationale Politik und Wirtschaft berichtet, konnten bestätigt werden. Insgesamt zeigt sich, dass die redaktionellen Strategien der Zeitungen in ihren Publikationen erkennbar sind.

Eingereicht von: Antonio Aleksic

Matrikelnummer: 200092

14.02.2025

Inhaltsverzeichnis

1. Einleitung 1

2. Daten und Methoden 2

3. Ergebnisse 3

4. Diskussion 5

Literaturverzeichnis i

Abbildungsverzeichnis

Abbildung 1: Top 3 Topics pro Zeitung 4

1. Einleitung

Zeitungen spielen eine zentrale Rolle in der öffentlichen Meinungsbildung und informieren über eine Vielzahl von gesellschaftlichen, politischen und wirtschaftlichen Themen. Dabei verfolgt jede Zeitung eine spezifische redaktionelle Strategie, die sich in der Auswahl und Gewichtung der publizierten Inhalte widerspiegelt.

Die Frankfurter Allgemeine Zeitung (FAZ) konzentriert sich vor allem auf Wirtschaft, Politik und Kultur (Frankfurter Allgemeine Zeitung, o. J.). Die Süddeutsche Zeitung (SZ) deckt ein breites Themenspektrum ab, darunter Politik, Wirtschaft, Feuilleton, Medien, Sport und Wissenschaft (Süddeutscher Verlag, o. J.). Die Welt wiederum konzentriert sich auf Politik, Wirtschaft, Finanzen, Kultur, Wissenschaft und internationale Berichterstattung (WELT, o. J.).

In der analytischen Forschung werden verschiedene Methoden eingesetzt, um thematische Schwerpunkte zu identifizieren. Eine häufig angewandte Technik ist die Themenmodellierung, die es ermöglicht, latente thematische Strukturen in großen Textkorpora zu erkennen. Dabei werden Algorithmen wie Latent Dirichlet Allocation (LDA) verwendet, um Dokumente aufgrund ihrer Wortverteilung bestimmten Themen zuzuordnen (Murel & Kavlakoglu, 2024).

Trotz klar definierter redaktioneller Strategien ist unklar, inwieweit sich diese tatsächlich in den veröffentlichten Artikeln widerspiegeln. Die zentrale Frage dieser Untersuchung lautet daher: Spiegeln sich die redaktionellen Strategien von FAZ, SZ und Welt in den veröffentlichten Artikeln wider?

Um diese Frage zu beantworten, wird ein datengetriebenes Verfahren angewendet, das auf der automatisierten Analyse der Startseiten dieser Zeitungen basiert. Mittels Web-Scraping wurden über einen Zeitraum von drei Monaten (November 2024 - Januar 2025) täglich die Startseiten der drei Zeitungen erfasst und in einer Datenbank gespeichert. Anschließend wurde mit Hilfe von LDA-Modellen eine Themenanalyse durchgeführt, um die zentralen Themen jeder Zeitung zu identifizieren und mit ihren redaktionellen Strategien abzugleichen.

Basierend auf den redaktionellen Ausrichtungen der Zeitungen wurden folgende Hypothesen aufgestellt:

- H1: Die FAZ wird überwiegend Themen aus den Bereichen Politik und Wirtschaft behandeln.
- H2: Die SZ wird ein breiteres Themenspektrum aufweisen, insbesondere mit Schwerpunkten in Gesellschaft, Kultur und Wissenschaft.
- H3: Die Welt wird sich stark auf internationale Politik, Wirtschaft und Technologie konzentrieren.

In den folgenden Kapiteln werden die Methodik der Datenerhebung und -auswertung sowie die gewonnenen Ergebnisse und deren Bedeutung für die publizistische Strategie der untersuchten Zeitungen beschrieben.

2. Daten und Methoden

Die Daten stammen aus einem Web-Scraping-Projekt, bei dem über einen längeren Zeitraum täglich die Startseiten mehrerer Zeitungen gesammelt wurden. Der Fokus lag dabei auf drei renommierten deutschen Publikationen:

- Frankfurter Allgemeine Zeitung (FAZ)
- Süddeutsche Zeitung (SZ)
- Die Welt

Der Untersuchungszeitraum erstreckte sich von November 2024 bis Januar 2025 und dauerte insgesamt drei Monate. In diesem Zeitraum wurden für jede Zeitung täglich die Startseiten gespeichert, was zu insgesamt 92 Startseiten pro Zeitung führte. Insgesamt wurden also 276 Startseiten gespeichert mit mehreren Artikeln.

Zur Extraktion der relevanten Artikel aus den HTML-Dateien wurde BeautifulSoup verwendet. Dabei wurden für jeden Artikel die folgenden Informationen extrahiert:

- Name der Zeitung
- Pfad der gespeicherten HTML-Datei
- Datum der Veröffentlichung
- Überschrift des Artikels
- Vorschautext des Artikels (falls vorhanden)

Die extrahierten Daten wurden in einer SQLite-Datenbank gespeichert. Der Datensatz umfasste insgesamt 45.143 Artikel.

Für die weitere Analyse wurde der Datensatz zunächst in einem DataFrame aufbereitet, wobei mehrere Bereinigungsschritte durchgeführt wurden. Zunächst wurden unvollständige Einträge entfernt, wobei alle Artikel ohne Überschrift ausgeschlossen wurden, was insgesamt 557 Artikel betraf. Anschließend wurden alle Wörter in Kleinbuchstaben umgewandelt, um die Modellverarbeitung zu optimieren. Es folgte die Entfernung von Sonderzeichen, wobei Zeichen wie „!“, „?“, „@“ und andere nicht-alphanumerische Symbole entfernt wurden. Schließlich wurden sogenannte Stoppwörter, d.h. häufig vorkommende, aber wenig informative Wörter wie „und“, „der“, „die“ und „das“ aus dem Datensatz entfernt.

Für die weitere Analyse wurde eine neue Spalte angelegt, die Titel und Text eines Artikels kombiniert. Diese Spalte dient als zentrale Grundlage für das Modell, das auf den vorverarbeiteten Textdaten basiert.

Durch diese Vorverarbeitung wurde eine strukturierte und qualitativ hochwertige Datenbasis geschaffen, die eine präzise Analyse ermöglicht.

3. Ergebnisse

Um die Hauptthemen innerhalb der Artikel zu identifizieren, wurde das Latent Dirichlet Allocation (LDA) Modell von sklearn verwendet. Für jede Zeitung wurde ein separates LDA-Modell trainiert.

Vor der Anwendung von LDA wurde der CountVectorizer verwendet. Dieser wandelt den Text in eine numerische Darstellung um. Dabei wird für jedes Wort gezählt, wie oft es in einem Artikel vorkommt. So kann das LDA-Modell anhand der Wortverteilung verschiedene Themen erkennen.

Nach dem Training des Modells wurden die Top-Wörter für jedes identifizierte Thema extrahiert und zusammengefasst. Anschließend wurden die Wahrscheinlichkeiten für jedes Thema berechnet. Um eine strukturierte Übersicht zu erhalten, wurden die Themen nach absteigender Wahrscheinlichkeit sortiert.

Die Analyse der Artikel ergab für jede Zeitung die zehn relevantesten Themen. Die Top-3-Themen pro Zeitung wurden in einer Grafik zusammengefasst, die die thematische Gewichtung je Publikation verdeutlicht (Abbildung 1). Daraus lassen sich folgende Beobachtungen ableiten:

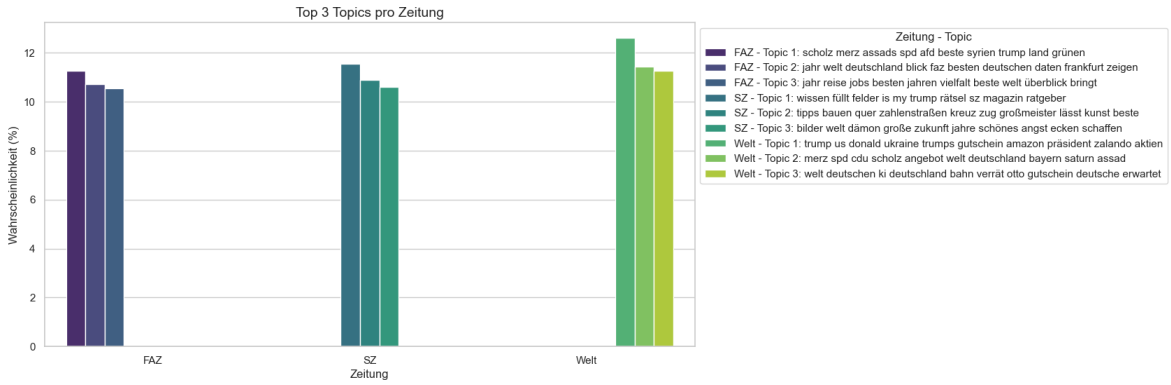

Abbildung 1: Top 3 Topics pro Zeitung

Quelle: Eigene Darstellung

FAZ legt besonderen Fokus auf politische Entwicklungen und wirtschaftliche Themen. SZ hebt gesellschaftliche und kulturelle Themen hervor. Welt berichtet stark über internationale Politik, insbesondere die USA und Technologie. Diese Ergebnisse zeigen, dass jede Zeitung ihre eigenen thematischen Schwerpunkte setzt, was Rückschlüsse auf die redaktionellen Ausrichtungen und Zielgruppen zulässt.

FAZ - Top 3 Themen

1. Politik und internationale Beziehungen (Scholz, Merz, Assad, Trump, SPD, AFD)
2. Jahresrückblick und Deutschland-Themen (Deutschland, FAZ, beste, Daten, Frankfurt)
3. Reise & Wirtschaft (Reisen, Jobs, Vielfalt, Überblick)

SZ - Top 3 Themen

1. Bildung & Wissen (Wissen, Rätsel, Intelligenz, Geschichte)
2. Gesellschaft & Kultur (Bilder, Angst, Zukunft, Kunst)
3. Gesundheit & Lebensart (Tipps, Gewicht, Lebensqualität)

Welt - Top 3 Themen

1. Internationale Politik & USA (Trump, USA, Ukraine, Präsident)
2. Deutschland & Innenpolitik (Merz, Scholz, CDU, SPD, Deutschland)
3. Technologie & Wirtschaft (KI, Deutsche Bahn, Aktien, Musk)

4. Diskussion

Die Ergebnisse der Themenmodellierung zeigen deutliche Unterschiede in den thematischen Schwerpunkten der untersuchten Zeitungen.

Die FAZ weist eine deutliche Dominanz der Themen Politik und Wirtschaft auf, was ihrer redaktionellen Strategie entspricht. Besonders häufig tauchen Begriffe wie Scholz, Merz, SPD, Deutschland, Wirtschaft und Daten auf, was bestätigt, dass sich diese Zeitung auf die Wirtschafts- und Politikberichterstattung konzentriert. Somit kann H1 bestätigt werden.

Die SZ weist eine größere Themenvielfalt auf. Neben Politik und Wirtschaft sind auch Kultur, Gesellschaft und Wissenschaft stark vertreten. Das Vorkommen der Begriffe Wissen, Feuilleton, Bilder, Kultur, Wissenschaft und Gesellschaft unterstützt die Hypothese, dass die SZ eine vielfältigere Berichterstattung verfolgt. H2 wird ebenfalls bestätigt.

Die Welt legt einen starken Fokus auf internationale Politik, Wirtschaft und Technologie, was sich in Themen wie Trump, Ukraine, Deutschland, Musk, Aktien und Handel widerspiegelt. Dies bestätigt ihren Fokus auf globale wirtschaftliche und politische Zusammenhänge. Auch H3 wird durch die Analyse bestätigt.

Obwohl die Ergebnisse die redaktionellen Strategien der Zeitungen weitgehend widerspiegeln, gibt es methodische Einschränkungen. Eine Einschränkung der Datenbasis besteht darin, dass die Analyse ausschließlich auf den Startseiten der

Zeitungen basiert. Es ist daher möglich, dass in anderen Bereichen der Websites andere thematische Schwerpunkte gesetzt werden, die in der Untersuchung nicht berücksichtigt wurden. Auch die Anwendung der LDA-Modellierung ist begrenzt, da diese Methode Schwierigkeiten bei der Erkennung kontextueller Bedeutungen hat. Themen können sich überschneiden oder falsch zugeordnet werden. Auch die Textbereinigung kann die Themenzuordnung beeinflussen, da die Vorverarbeitung der Daten, wie z.B. das Entfernen von Stoppwörtern, relevante Kontextinformationen reduzieren kann. Ein weiterer limitierender Faktor ist der Analysezeitraum, der nur drei Monate umfasst. Dadurch können saisonale oder kurzfristige Ereignisse die Ergebnisse beeinflussen, was die Generalisierbarkeit der Ergebnisse einschränkt.

Die Analyse zeigt, dass sich die redaktionellen Strategien von FAZ, SZ und Welt in den veröffentlichten Artikeln widerspiegeln. Die Themenmodellierung bestätigt die erwarteten Schwerpunkte der einzelnen Zeitungen, auch wenn methodische Einschränkungen berücksichtigt werden müssen. Zukünftige Forschung könnte die Analyse auf längere Zeiträume ausdehnen oder weitere Zeitungen einbeziehen, um ein noch genaueres Bild der Berichterstattung zu erhalten.

Literaturverzeichnis

Frankfurter Allgemeine Zeitung (o. J.). Die FAZ – Profil der Frankfurter Allgemeinen Zeitung. Verfügbar unter: https://www.frankfurterallgemeine.de/die-faz (Zugriff am: 14.02.2025).

Murel, J. & Kavlakoglu, E. (2024). Was ist Topic Modeling?. IBM. Verfügbar unter: https://www.ibm.com/de-de/topics/topic-modeling (Zugriff am: 14.02.2025

Süddeutscher Verlag (o. J.). Süddeutsche Zeitung – Profil. Verfügbar unter: https://www.sueddeutscher-verlag.de/sueddeutsche-zeitung (Zugriff am: 14.02.2025).

WELT (o. J.). Impressum. Verfügbar unter:

https://www.welt.de/services/article104636888/Impressum.html (Zugriff am: 14.02.2025).

Eigenständigkeitserklärung für schriftliche Prüfungsleistungen an der Digital Business University of Applied Sciences

Name	Aleksic
Vorname	Antonio
Matrikelnummer	200092
Modultitel	ADS-01: Tools der Softwareentwicklung und Online-Daten
Thema der Prüfungsleistung	Themenmodellierung mittels Latent Dirichlet Allocation auf Zeitungshomepages
Prüfer:in	Prof. Dr. Marcel Hebing
Datum	14.02.2025

Ich trage die Verantwortung für die Qualität des Textes sowie die Auswahl aller Inhalte und habe sichergestellt, dass Informationen und Argumente mit geeigneten wissenschaftlichen Quellen belegt bzw. gestützt werden. Die aus fremden Quellen direkt oder indirekt übernommenen Texte, Gedankengänge, Konzepte, Grafiken usw. in meinen Ausführungen habe ich als solche eindeutig gekennzeichnet und mit vollständigen Verweisen auf die jeweilige Quelle versehen. Alle weiteren Inhalte dieser Arbeit (Textteile, Abbildungen, Tabellen etc.) ohne entsprechende Verweise stammen im urheberrechtlichen Sinn von mir.

☒ Hiermit erkläre ich, dass ich die vorliegende Prüfungsleistung selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle sinngemäß und wörtlich übernommenen Textstellen aus fremden Quellen wurden kenntlich gemacht.

☒ Die vorliegende Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Erklärung zu (gen)KI-Tools

☒ **Verwendung von (gen)KI-Tools**

Ich versichere, dass ich mich (gen)KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Ich verantworte die Übernahme jeglicher von mir verwendeter Textpassagen vollumfänglich selbst. In der „Übersicht verwendeter (gen)KI-Tools" habe ich sämtliche eingesetzte (gen)KI-Tools, deren Einsatzform sowie die jeweils betroffenen Teile der Arbeit einzeln aufgeführt. Ich versichere, dass ich keine (gen)KI-Tools verwendet habe, deren Nutzung der Prüfer bzw. die Prüferin explizit schriftlich ausgeschlossen hat.

Hinweis: Sofern die zuständigen Prüfenden bis zum Zeitpunkt der Ausgabe der Aufgabenstellung konkrete (gen)KI-Tools ausdrücklich als nicht anzeige-/kennzeichnungspflichtig benannt haben, müssen diese nicht aufgeführt werden.

Ich erkläre weiterhin, dass ich mich aktiv über die Leistungsfähigkeit und Beschränkungen der unten genannten (gen)KI-Tools informiert habe und überprüft habe, dass die mithilfe der genannten (gen)KI-Tools generierten und von mir übernommenen Inhalte faktisch richtig sind.

☒ **Verbot bzw. Nicht-Nutzung von (gen)KI-Tools**

Ich versichere, dass ich die hier vorliegende Arbeit vollständig eigenständig formuliert habe und keine (gen)KI-Tools verwendet habe.

Mir ist bekannt, dass ein Verstoß gegen die genannten Punkte prüfungsrechtliche Konsequenzen haben und insbesondere dazu führen kann, dass die Prüfungsleistung mit "nicht ausreichend" bzw. die Studienleistung mit „nicht bestanden" bewertet wird und bei mehrfachem oder schwerwiegendem Täuschungsversuch eine Exmatrikulation erfolgen kann.

Übersicht verwendeter (gen)KI-Tools

Die (gen)KI-Tools habe ich, wie im Folgenden dargestellt, eingesetzt.

(gen)KI-Tool	**Einsatzform**	**Betroffene Teile der Arbeit**
Deepl	Textverbesserung	

Ort	Berlin
Datum	14.02.205

Unterschrift Student:in