WS 2023/24

Seminar Ökonometrie: Text Mining

Prof. Dr. Peter Winker
e–mail: Peter.Winker@wirtschaft.uni-giessen.de Tel.: 0641/99-22-640

Elena Tönjes
e–mail: Elena.Toenjes@wirtschaft.uni-giessen.de Tel.: 0641/99-22-643

Das Seminar behandelt die Anwendung von computerlinguistischen Techniken für ökonomische Fragestellungen. Es werden unterschiedliche Methoden erarbeitet, die in diesem Anwendungsgebiet in der Praxis relevant und/oder Gegenstand aktueller Forschungsanstrengungen sind. Neben dem Verständnis der methodischen Grundlagen wird auch die empirische Umsetzung der Methoden kritisch diskutiert. Dazu sollen die Seminarteilnehmer weitere empirische Evidenz aus der Literatur und/oder eigenen empirischen Arbeiten beisteuern. Ziel des Seminars ist es, die TeilnehmerInnen zu einer selbstständigen Auseinandersetzung mit Methoden und Anwendungen aus dem Bereich des Natural Language Processing zu befähigen, die Lektüre und Auswertung wissenschaftlicher Beiträge einzuüben und die Darstellung und den Vortrag wissenschaftlicher Resultate zu trainieren. Fortgeschrittene Kenntnisse in Statistik sowie Programmierkenntnisse werden empfohlen.

Themenübersicht

Thema 1:	Der Wert von Textdaten im wirtschaftlichen Bereich
Thema 2:	Der Wert von Textdaten im wissenschaftlichen Bereich
Thema 3:	Datensätze und Aufbereitung von Text Daten
Thema 4:	Text Repräsentationen (*2Vec)
Thema 5:	Wörterbuch basierte Sentiment Analyse
Thema 6:	Machine-Learning basierte Sentiment Analyse
Thema 7:	Textklassifizierung
Thema 8:	Topic Modeling
Thema 9:	Effektive Beurteilung von 'unsupervised' NLP-Modellen: Methoden und Herausforderungen
Thema 10:	Named Entity Recognition
Thema 11:	Text Summarization
Thema 12:	Text-basierte Indikatoren
Thema 13:	Text Mining im Finanzbereich
Thema 14:	Large Language Models: Ernie, Bert, Grover, XLNet, GPT-2/3/4
Thema 15:	Die Evolution der Interaktion: Auswirkungen von AI-Textgenerierung auf Wissenschaft, Forschung und Arbeitsprofile

Basisliteratur

- Zusätzliche Literatur wird für die einzelnen Seminarthemen bereitgestellt. Die angegebene Literatur dient jedoch lediglich als Ausgangspunkt und sollte durch weitere Recherchen mit aktueller Literatur ergänzt werden.

Voraussetzungen

- regelmäßige aktive Teilnahme an den Seminarsitzungen
- Anfertigung einer mit mindestens "ausreichend" (4,0 bzw. 5 Punkten) bewerteten Seminararbeit (Gewichtung 2/3)
- Vortrag der Arbeit im Rahmen einer Seminarsitzung sowie kurzes Koreferat zu einem anderen Seminarthema (Gewichtung 1/3)
- **Gute Grundlagen in Statistik, Mathematik und Ökonometrie!**

Weitere Informationen

ECTS Punkte:	6
Sem.-Wochenstunden:	2
Umfang der Arbeit:	ca. 10 Seiten (+/- 10 %, Text inkl. Abbildungen und Tabellen)
Abgabe vorläufige Arbeit	eine Woche vor der jeweiligen Präsentation
Abgabe endgültige Version	bis spätestens 23.02.2024.
Sprache	Englisch oder Deutsch
Dauer der Präsentation	max. 30 Minuten

Thema 1: Der Wert von Textdaten im wirtschaftlichen Bereich

Textdaten gewinnen immer mehr an Wert und Relevanz in einer Vielzahl an Bereichen. Die Themen 1 und 2 sollen zunächst ein Gefühl für die Relevanz von Texten als Daten vermitteln. Dieses Thema soll den Wert von Text als Daten im Bereich der Wirtschaft beleuchten und die zentrale Frage, die in dieser Seminararbeit beantwortet werden soll, lautet: Welche Textdaten werden im wirtschaftlichen Bereich eingesetzt und wie lässt sich ihr Wert bestimmen? Konkretere Fragen wären: Wie wertvoll sind z.B. Bewertungen für ein Hotel, Restaurant oder auch den Kunden? Welchen monetären Wert haben solche Bewertungen? Diese oder ähnliche Fragen sollen in Form einer Literaturübersicht beantwortet werden.

Einstiegsliteratur:

- Matthew Gentzkow, Bryan T Kelly, and Matt Taddy. Text as data. Working Paper 23276, National Bureau of Economic Research, March 2017
- Patricia Cleary, Kristen Garlock, Denise Novak, Ethan Pullman, and Sanjeet Mann. Text mining 101: What you should know. *The Serials Librarian*, 72(1-4):156–159, 2017
- Gerhard Heyer, Uwe Quasthoff, and Thomas Wittig. Text mining: Wissensrohstoff text. *W3l, Herdecke*, 18, 2006

Thema 2: Der Wert von Textdaten im wissenschaftlichen Bereich

Textdaten gewinnen immer mehr an Wert und Relevanz in einer Vielzahl an Bereichen. Thema 1 und 2 sollen zunächst ein Gefühl für die Relevanz von Texten als Daten vermitteln. Dieses Thema soll den Wert von Text als Daten im Bereich Forschung mit besonderem Fokus auf die Wirtschaftswissenschaften beleuchten und die zentrale Frage, die in dieser Seminararbeit beantwortet werden soll, lautet: Welche Textdaten werden für wissenschaftliche Fragestellungen im Bereich der Wirtschaftswissenschaften eingesetzt? Konkretere Fragen in diesem Kontext wären: Wie wertvoll sind Unternehmensdaten in Form von Text auf deren Websites für die Forschung? Wie relevant sind z.B. Zeitungsartikel oder Unternehmensberichte? Diese Fragen sollen in Form einer Literaturübersicht beantwortet werden.

Einstiegsliteratur:

- Matthew Gentzkow, Bryan T Kelly, and Matt Taddy. Text as data. Working Paper 23276, National Bureau of Economic Research, March 2017
- Patricia Cleary, Kristen Garlock, Denise Novak, Ethan Pullman, and Sanjeet Mann. Text mining 101: What you should know. *The Serials Librarian*, 72(1-4):156–159, 2017
- Gerhard Heyer, Uwe Quasthoff, and Thomas Wittig. Text mining: Wissensrohstoff text. *W3l, Herdecke*, 18, 2006

Thema 3: Datensätze und Aufbereitung von Text Daten

Das Sammeln von Daten ist der erste Schritt bei der datengetriebenen Analyse einer gegebenen Problemstellung. Organisationen stellen häufig öffentliche Programmierschnittstellen für den Zugriff auf ihre Daten bereit, bspw. die Twitter API oder die NY Times API. Auch gibt es eine Fülle von frei verfügbaren Datensätzen, bspw. auf Github, Kaggle oder Reddit. Diese Daten sind in ihrer rohen Form häufig nicht direkt für ökonomische Analysen nutzbar. Gegenstand der Seminararbeit sollte die Darstellung aktueller Standards in der Datenaufbereitung sein, bspw. Stemming, Lemmatization, Stopwords filtern, Popularitäts-basiertes Vorfiltern von Wörtern etc.

Einstiegsliteratur:

- Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. *Text Preprocessing*, pages 45–59. Springer International Publishing, Cham, 2019
- S. Vijayarani, M. Ilamathi, and Ms. Nithya. Preprocessing techniques for text mining-an overview dr. 2015

Thema 4: Text Repräsentationen

Text-Daten sind typischerweise hoch-dimensional und unstrukturiert. Die Art der Aufbereitung und Darstellung von Text-Daten für den Gebrauch mit Computern ist von entscheidender Bedeutung für den Erfolg der anschließenden Analysen. Traditionelle Darstellungen, bspw. die Repräsentation einzelner Wörter als Indizes in einem Wörterbuch, sind schnell und einfach zu reproduzieren. Aktuelle Forschungsanstrengungen zeigen, dass komplexere Methoden, bspw. Word2Vec oder fastText, die Ergebnisse jedoch deutlich verbessern können. Gegenstand der Seminararbeit sollte ein Vergleich verschiedener Methoden, bspw. BoW / Word2Vec / GloVe / fastText sein, oder die Anwendung selbiger, oder die detaillierte Darstellung einer der neueren Methoden.

Einstiegsliteratur:

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016

- David Lenz and Peter Winker. Measuring the diffusion of innovations with paragraph vector topic models. *PloS one*, 15(1):e0226685, 2020

Thema 5: Wörterbuch basierte Sentiment Analyse

Im Allgemeinen zielt die Stimmungsanalyse darauf ab, die Einstellung eines Sprechers, Schreibers oder eines anderen Subjekts in Bezug auf ein Thema oder die gesamte kontextuelle Polarität oder emotionale Reaktion auf ein Dokument, eine Interaktion oder ein Ereignis zu bestimmen. Die Sentimentanalyse von Freitextdokumenten ist eine gängige Aufgabe im Bereich Text Mining. In der klassischen Sentiment-Analyse werden Texten vordefinierte Sentiment-Labels wie "positiv" oder "negativ" zugeordnet. Texte (hier als Dokumente bezeichnet) können Rezensionen über Produkte oder Filme, Artikel usw. sein. Traditionelle Methoden benutzen Wörterbücher mit positiven und negativen Wörtern, um von der Häufigkeit der jeweiligen Wortkategorie auf den Sentiment eines Dokuments zu schließen. Inhalt der Seminararbeit sollte die Darstellung Wörterbuch basierter Methoden sein, oder die Anwendung einer solchen Methode.

Einstiegsliteratur:

- Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16 – 32, 2018
- Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012
- Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. *A practical guide to sentiment analysis*, volume 5. Springer, 2017
- Christina E Bannier, Thomas Pauls, and Andreas Walter. Content analysis of business communication: Introducing a german dictionary. *Journal of Business Economics*, 2018

Thema 6: Machine-learning basierte Sentiment Analyse

Im Allgemeinen zielt die Stimmungsanalyse darauf ab, die Einstellung eines Sprechers, Schreibers oder eines anderen Subjekts in Bezug auf ein Thema oder die gesamte kontextuelle Polarität oder emotionale Reaktion auf ein Dokument, eine Interaktion oder ein Ereignis zu bestimmen. Die Sentimentanalyse von Freitextdokumenten ist eine gängige Aufgabe im Bereich Text Mining. In der klassischen Sentiment-Analyse werden Texten vordefinierte Sentiment-Labels wie "positiv" oder "negativ" zugeordnet. Texte (hier als Dokumente bezeichnet) können Rezensionen über Produkte oder Filme, Artikel usw. sein. Aktuelle Methoden benutzen Machine Learning, um auf den Sentiment eines Doukments zu schließen. Inhalt der Seminararbeit könnte die Darstellung Machine Learning-basierter Methoden sein, oder die Anwendung einer machine

learning Methode zur Sentiment Bestimmung.

Einstiegsliteratur:

- Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16 – 32, 2018
- Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012
- Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. *A practical guide to sentiment analysis*, volume 5. Springer, 2017

Thema 7: Textklassifizierung

Die Klassifizierung von Textdokumenten nach Kategorien oder Themen ist ein wichtiger Bestandteil vieler Textverarbeitungssysteme. Die Aufgabe der Textklassifizierung besteht darin, Dokumente in vordefinierte Themen wie Wirtschaft, Politik und Sport einzuteilen. Zum Beispiel ist die automatische Zuweisung mehrerer klinischer Codes (fachspezifische medizinische Ausdrücke) zu klinischem Freitext ein typisches Problem der Textklassifikation mit mehreren Themen. Spam-Erkennung ist ein weiteres Beispiel für Analysen dieser Kategorie. Inhalt der Seminararbeit könnte die Darstellung verschiedener Anwendungen und Methoden der Textklassifizierung sein, oder eine eigene Untersuchung, bei der Texte in Kategorien unterteilt werden.

Einstiegsliteratur:

- Ashok Srivastava and Mehran Sahami. *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC, Boca Raton, 2009, (verfügbar an der Professur) Kap. 3, 7
- Ronny Luss and Alexandre d'Aspremont. Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6):999–1012, 2015
- Alan S Abrahams, Weiguo Fan, G Alan Wang, Zhongju Zhang, and Jian Jiao. An integrated text analytic framework for product defect discovery. *Production and Operations Management*, 24(6):975–990, 2015

Thema 8: Topic Modeling

Beim maschinellen Lernen und bei der Verarbeitung natürlicher Sprache ist ein Topic Modell eine Art statistisches Modell zum Auffinden der abstrakten "Themen", die in einer Sammlung von Dokumenten vorkommen. Topic Modelling ist ein häufig verwendetes Text-Mining-Werkzeug zum Auffinden verborgener semantischer Strukturen in einem Textkörper. Inhalt der Seminararbeit sollte die detaillierte Darstellung einer der Methoden, bspw. LDA, sein, oder ein Vergleich verschiedener Topic Modelle, oder die Anwendung einer Methode zur Identifikation latenter Topics in einem Corpus.

Einstiegsliteratur:

- Ashok Srivastava and Mehran Sahami. *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC, Boca Raton, 2009 (verfügbar an der Professur) Kap. 4
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003
- Ryohei Hisano, Didier Sornette, Takayuki Mizuno, Takaaki Ohnishi, and Tsutomu Watanabe. High quality topic extraction from business news explains abnormal financial market volatility. *PLOS ONE*, 8(6):1–12, 06 2013
- Takayuki Mizuno, Takaaki Ohnishi, and Tsutomu Watanabe. Novel and topical business news and their impact on stock market activity. *EPJ Data Science*, 6(1):26, Oct 2017
- David Lenz and Peter Winker. Measuring the diffusion of innovations with paragraph vector topic models. *PloS one*, 15(1):e0226685, 2020

Thema 9: Effektive Beurteilung von 'unsupervised' NLP-Modellen: Methoden und Herausforderungen

Die Evaluierung von 'unsupervised' NLP-Modellen unterscheidet sich maßgeblich von der Evaluierung von 'supervised' Modellen und bringt gewisse Herausforderungen mit sich. In der Arbeit sollen aktuelle Methoden dargestellt werden, um die Leistung dieser Modelle objektiv zu messen, darunter Techniken zur Beurteilung von Textkohärenz, semantischem Reichtum und Transferfähigkeit. Zudem sollen die Herausforderungen diskutieret werden, die bei der Evaluierung auftreten können, wie die Schwierigkeit, angemessene Ground-Truth-Daten zu erstellen, da diese Modelle oft keine klaren Antwortvorgaben haben. Des Weiteren sollen komplexen Fragen beleuchtet werden, die mit der Bewertung von Textqualität und dem Umgang mit subjektiven Kriterien verbunden sind.

Einstiegsliteratur:

- Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation

methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009

- Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, page 102131, 2022

Thema 10: Named Entity Recognition

Text Daten beinhalten eine Fülle von Informationen. Für eine Reihe von Analysen ist es relevant zu wissen, welche Entitäten in einem Text von Bedeutung sind. Entitäten sind namentlich benennbare reale Personen oder Konstrukte wie bspw. Unternehmen, Organisationen oder Staaten. Für ökonomische Anwendungen ist es häufig relevant zu wissen, welche Entitäten in einer Analyse von Bedeutung sind. Gegenstand der Seminararbeit könnte eine Darstellung aktueller Methoden und Anwendungen von Named Entity Recognition Methoden sein.

Einstiegsliteratur:

- Behrang Mohit. *Named Entity Recognition*, pages 221–245. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014

- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360, 2016

Thema 11: Text Summarization

Automatische Zusammenfassung ist der Prozess der Verkürzung eines Textdokuments mit Software, um eine Zusammenfassung mit den Hauptpunkten des Originaldokuments zu erstellen. Technologien, die eine kohärente Zusammenfassung machen können, berücksichtigen Aspekte wie Länge, Schreibstil und Syntax. Inhalt der Seminararbeit könnte die Darstellung aktueller Forschungsanstrengungen oder die Implementation eines Text Summarization Systems sein.

Literatur:

- Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. Text summarization techniques: A brief survey. *CoRR*, abs/1707.02268, 2017

- Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. Automatic labelling of topic models learned from twitter by summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 618–624, 2014

Thema 12: Text-basierte Indikatoren

Die Messung (& Prognose) mikro- und makro-ökonomischer Größen wie BIP, Zins oder Wertpapierkurse ist von großer Relevanz, wichtige Merkmale sind eine schnelle Verfügbarkeit und hohe Treffsicherheit. Traditionelle Indikatoren haben verschiedene Probleme, bspw. langsame Verfügbarkeit oder große Prognosefehler bei sich verändernden Marktbedingungen. Text Daten, bspw. Zeitungsartikel, beinhalten eine große Menge relevanter Informationen, die Marktakteure in Echtzeit zur Steuerung ihrer wirtschaftlichen Aktivitäten nutzen. Die Konstruktion Text-basierter Indikatoren ist daher Gegenstand aktueller Forschungen. Inhalt der Seminararbeit könnte die Darstellung / Gegenüberstellung verschiedener Text-basierter Indikatoren sein, oder die Konstruktion eines Text-basierten Indikators.

Einstiegsliteratur:

- Abby Levenberg, Stephen Pulman, Karo Moilanen, Edwin Simpson, and Stephen Roberts. Predicting economic indicators from Web text using sentiment composition. *International Journal of Computer and Communication Engineering*, (2), 2014
- Jochen Lüdering and Peter Winker. Forward or backward looking? The economic discourse and the observed reality. *Jahrbücher für Nationalökonomie und Statistik*, 236(4):483–515, 2016
- Samuel Rönnqvist and Peter Sarlin. Bank distress in the news: Describing events through deep learning. *Neurocomputing*, 264:57 – 70, 2017. Machine learning in finance
- Leif Anders Thorsrud. Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, (online):1–35, 2018
- Peter Tillmann and Andreas Walter. ECB vs bundesbank: Diverging tones and policy effectiveness. *MAGKS Discussion Paper*, 20, 2018

Thema 13: Text Mining im Finanzbereich

Text-Daten enthalten Informationen, die für viele Bereiche der Wirtschaft nützlich sein können, so auch für den Finanzbereich. Texte können Aufschluss über Verhaltensweisen, Stimmung, Meinungen und Beziehungen geben, welche sich nicht unbedingt in Zahlen widerspiegeln. Solche Informationen können beispielsweise im Bereich Behavioral Finance, bei der Vorhersage von Marktbewegungen, im Bereich Accounting, im Risikomanagement oder auch im Portfolio Management nützlich sein. Inhalt der Seminararbeit könnte die Darstellung von Text Mining in einem Finanzbereich oder die Anwendung einer NLP-Methode im Finanzbereich sein.

Einstiegsliteratur:

- Sanjiv Ranjan Das et al. Text and context: Language analytics in finance. *Foundations*

and Trends® in Finance, 8(3):145–261, 2014

- Gautam Mitra and Leela Mitra. *The handbook of news analytics in finance*, volume 596. John Wiley & Sons, 2011
- Ronny Luss and Alexandre d'Aspremont. Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6):999–1012, 2015
- Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19, 2009
- Dania Eugenidis, David Lenz, Christoph Leser, Frauke Schleer-van Gellecom, and Peter Winker. Text-mining basierte analyse der Kapitalmarktreaktionen auf Ad-hoc-Mitteilungen. *CORPORATE FINANCE*, 2020

Thema 14: Large Language Models: Ernie, Bert, Grover, XLNet, GPT-2/3/4

Transfer Learning, Modelle, die auf riesige Datenmengen vortrainiert wurden, um ein generelles Verständnis für Sprache zu entwickeln, und dann auf kleinere Task-spezifische Datensätze nachtrainiert werden, sind momentan die am Besten funktionierenden Methoden im NLP Research. Sie zeigen potentielle Anwendungsmöglichkeiten für ökonomische Fragestellungen auf oder wenden eine SOTA Technik an um ein Problem zu lösen. Inhalt der Seminararbeit ist, den aktuellen Stand der Modelle im Bereich Text Mining darzustellen.

Einstiegsliteratur:

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. *CoRR*, abs/1905.07129, 2019
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *CoRR*, abs/1905.12616, 2019
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019

- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020

- Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020

- Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models, 2023

Thema 15: Die Evolution der Interaktion: Auswirkungen von AI-Textgenerierung auf Wissenschaft, Forschung und Arbeitsprofile

Im Fokus dieses Seminarabschnitts steht der weitreichende Einfluss von AI-Textgenerierungsmodellen wie ChatGPT. Wir werden die konkreten Auswirkungen auf wissenschaftliche und forschungsbasierte Prozesse untersuchen, von der Hypothesenbildung bis zur Ergebnispräsentation. Darüber hinaus werfen wir einen Blick auf Berufsfelder, in denen die Automatisierung von Texterstellung möglicherweise zu Veränderungen führen könnte, und diskutieren potenzielle Ansätze zur Neugestaltung dieser Arbeitsprofile. Dieser Abschnitt ermöglicht es uns, Chancen und Herausforderungen dieser Technologie in Bezug auf kreative, wissenschaftliche und berufliche Aktivitäten zu erkunden.

Einstiegsliteratur:

- Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642, 2023

- Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410, 2023

- Xiaoming Zhai. Chatgpt user experience: Implications for education. *Available at SSRN 4312418*, 2022

- Ali Zarifhonarvar. Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Available at SSRN 4350925*, 2023

References

[1] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, page 102131, 2022.

[2] Alan S Abrahams, Weiguo Fan, G Alan Wang, Zhongju Zhang, and Jian Jiao. An integrated text analytic framework for product defect discovery. *Production and Operations Management*, 24(6):975–990, 2015.

[3] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. Text summarization techniques: A brief survey. *CoRR*, abs/1707.02268, 2017.

[4] Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. *Text Preprocessing*, pages 45–59. Springer International Publishing, Cham, 2019.

[5] Christina E Bannier, Thomas Pauls, and Andreas Walter. Content analysis of business communication: Introducing a german dictionary. *Journal of Business Economics*, 2018.

[6] Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. Automatic labelling of topic models learned from twitter by summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 618–624, 2014.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[8] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[9] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. *A practical guide to sentiment analysis*, volume 5. Springer, 2017.

[10] Patricia Cleary, Kristen Garlock, Denise Novak, Ethan Pullman, and Sanjeet Mann. Text mining 101: What you should know. *The Serials Librarian*, 72(1-4):156–159, 2017.

[11] Sanjiv Ranjan Das et al. Text and context: Language analytics in finance. *Foundations and Trends® in Finance*, 8(3):145–261, 2014.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[13] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642, 2023.

[14] Dania Eugenidis, David Lenz, Christoph Leser, Frauke Schleer-van Gellecom, and Peter Winker. Text-mining basierte analyse der Kapitalmarktreaktionen auf Ad-hoc-Mitteilungen. *CORPORATE FINANCE*, 2020.

[15] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[16] Matthew Gentzkow, Bryan T Kelly, and Matt Taddy. Text as data. Working Paper 23276, National Bureau of Economic Research, March 2017.

[17] Gerhard Heyer, Uwe Quasthoff, and Thomas Wittig. Text mining: Wissensrohstoff text. *W3l, Herdecke*, 18, 2006.

[18] Ryohei Hisano, Didier Sornette, Takayuki Mizuno, Takaaki Ohnishi, and Tsutomu Watanabe. High quality topic extraction from business news explains abnormal financial market volatility. *PLOS ONE*, 8(6):1–12, 06 2013.

[19] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[20] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360, 2016.

[21] David Lenz and Peter Winker. Measuring the diffusion of innovations with paragraph vector topic models. *PloS one*, 15(1):e0226685, 2020.

[22] Abby Levenberg, Stephen Pulman, Karo Moilanen, Edwin Simpson, and Stephen Roberts. Predicting economic indicators from Web text using sentiment composition. *International Journal of Computer and Communication Engineering*, (2), 2014.

[23] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[24] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models, 2023.

[25] Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410, 2023.

[26] Jochen Lüdering and Peter Winker. Forward or backward looking? The economic discourse and the observed reality. *Jahrbücher für Nationalökonomie und Statistik*, 236(4):483–515, 2016.

[27] Ronny Luss and Alexandre d'Aspremont. Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6):999–1012, 2015.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[30] Gautam Mitra and Leela Mitra. *The handbook of news analytics in finance*, volume 596. John Wiley & Sons, 2011.

[31] Takayuki Mizuno, Takaaki Ohnishi, and Tsutomu Watanabe. Novel and topical business news and their impact on stock market activity. *EPJ Data Science*, 6(1):26, Oct 2017.

[32] Behrang Mohit. *Named Entity Recognition*, pages 221–245. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[33] Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16 – 32, 2018.

[34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[35] Samuel Rönnqvist and Peter Sarlin. Bank distress in the news: Describing events through deep learning. *Neurocomputing*, 264:57 – 70, 2017. Machine learning in finance.

[36] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19, 2009.

[37] Ashok Srivastava and Mehran Sahami. *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC, Boca Raton, 2009.

[38] Leif Anders Thorsrud. Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, (online):1–35, 2018.

[39] Peter Tillmann and Andreas Walter. ECB vs bundesbank: Diverging tones and policy effectiveness. *MAGKS Discussion Paper*, 20, 2018.

[40] S. Vijayarani, M. Ilamathi, and Ms. Nithya. Preprocessing techniques for text mining-an overview dr. 2015.

[41] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.

[42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019.

[43] Ali Zarifhonarvar. Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Available at SSRN 4350925*, 2023.

[44] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *CoRR*, abs/1905.12616, 2019.

[45] Xiaoming Zhai. Chatgpt user experience: Implications for education. *Available at SSRN 4312418*, 2022.

[46] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. *CoRR*, abs/1905.07129, 2019.