Exam literature

Core material:

- All slides, handouts, and tutorials

Methodological textbook and articles:

- Van Atteveldt, W., & Trilling, D. & Arcila, C. (2021). Computational Analysis of Communication: A practical introduction to the analysis of text, networks, and images with code examples in Python and R. Wiley. https://cssbook.net/ (Chapters 1-11)
- Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. Communication Methods and Measures, 11(4), 245-265. https://doi.org/10.1080/19312458.2017.1387238
- Introduction to Computational Communication Science: Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. Communication Methods and Measures, 12(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Example articles referenced in the slides:

Note: You do not need to be able to reproduce the content of these articles or know the results by heart. We do expect you to understand how computational methods were used in these articles and what conclusions can be drawn from these methods and their results.

Week 2:

- Heidenreich, T., Eberl, J.-M., Lind, F. & Boomgaarden, H. (2020). Political migration discourses on social media: a comparative perspective on visibility and sentiment across political Facebook accounts in Europe. Journal of Ethnic and Migration Studies, (46)7, 1261-1280, https://doi.org/10.1080/1369183X.2019.1665990
- Mellado, C., Hallin, D., Cárcamo, L. Alfaro, R. … & Ramos, A. (2021) Sourcing Pandemic News: A Cross-National Computational Analysis of Mainstream Media Coverage of COVID-19 on Facebook, Twitter, and Instagram. Digital Journalism, 9(9), 1271-1295, https://doi.org/10.1080/21670811.2021.1942114

Week 3:

- Su, L. Y. F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. New Media & Society, 20(10), 3678-3699. https://doi.org/10.1177/1461444818757205
- de León, E., Vermeer, S., & Trilling, D. (2023). Electoral news sharing: A study of changes in news coverage and Facebook sharing behaviour during the 2018 Mexican elections. Information, Communication & Society, 26(6), 1193-1209. https://doi.org/10.1080/1369118X.2021.1994629

Week 4:

- Balluff, P., Eberl, J-M., Oberhänsli, S. J., Bernhard-Harrer, J., & Huber, M. (2024). The Austrian Political Advertisement Scandal: Patterns of "Journalism for Sale". International Journal of Press/Politics. https://doi.org/10.1177/19401612241285672

Mock exam questions (answer key on last page)

Mock questions week 1:

1.1 Computational Social Science – definition

Computational social science uses digital trace data because…

A. Digital trace data do not require preprocessing.
B. Digital trace data are always representative of the population.
C. Digital trace data can capture real-world behavior at large scale.
D. Digital trace data eliminate the need for theory-based research.

1.2 Computational Social Science: Why now?

Which factor best explains the current rise of computational approaches?

A. More digital behaviors are logged and stored automatically.
B. Traditional methods like surveys no longer produce valid data.
C. Text data became obsolete compared to image data.
D. Humans now rarely interact offline.

1.3 The Text Classification Pipeline: Feature engineering

What happens in the *feature engineering* step of the classification pipeline?

A. Assigning class labels based on model outputs
B. Evaluating predictions against a gold standard
C. Collecting documents from external sources
D. Turning texts into a numerical representation

1.4 Ethical concerns of AI

A major ethical concern in computational research that uses third-party AI systems (e.g., OpenAI, Google) is that…

A. AI algorithms cannot be used for large-scale analysis.
B. Personal data may be exposed to the company providing the software.
C. Proprietary systems are less accurate than open-source tools.
D. Data must be translated into another language before use.

Mock questions week 2:

2.1 Preprocessing

Which of the following is not typically done in preprocessing?

A. Tokenization
B. Annotation
C. Frequency trimming
D. Stop-word removal

2.2 Document-term matrix

A document-term matrix (DTM; also called a Document-Feature Matrix or DFM) …

A. Stores frequencies of words for each document.
B. Stores similarity scores between topics and documents.
C. Represents the syntactic structure of each sentence.
D. Contains metadata features about documents.

2.3 Rule-based methods

Why are dictionary approaches deterministic?

A. They produce the same output each time it is applied.
B. They automatically learn new word associations.
C. They rely on random initialization.
D. They use probabilistic inference.

2.4. Confusion matrix interpretation

If a dictionary misclassifies many neutral comments as toxic, but almost never misclassifies toxic comments as neutral, it has…

A. Low precision but high recall
B. High precision but low recall
C. High F-score but low recall
D. High F-score but low precision

2.5. Bag-of-words limitation

A key weakness of bag-of-words approaches is that they…

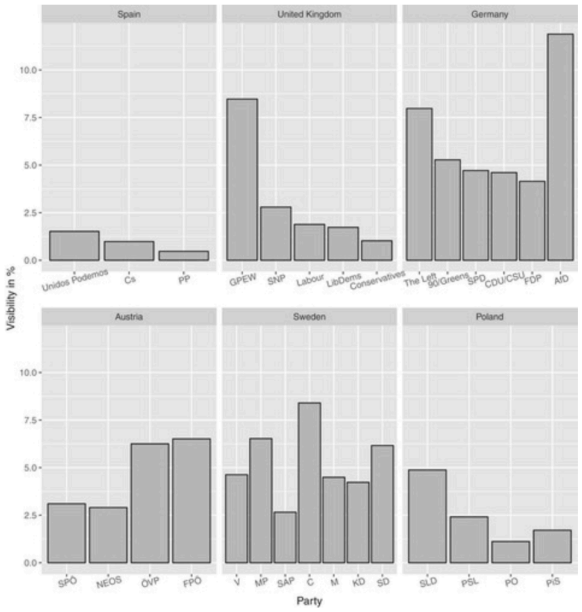
A. Cannot handle large corpora.
B. Prevent the use of machine-learning models.
C. Remove punctuation automatically.
D. Ignore word order and context.

2.6 Dictionary analysis of migration

Heidenreich et al. (2019) use a dictionary to measure mentions of migration in political parties' social media posts. They present the following figure:

Figure 1. Migration-related status posts as a proportion of the total.

Note: Data are percentages and shown grouped by country. Parties are shown left to right in order of the Chapel Hill left-right ideological score. For an overview see Table A1 in the Appendix.



What do the authors conclude from this figure?

A. Right-wing parties consistently post the most about migration in every country.
B. Migration visibility shows no ideological pattern and appears random across actors.
C. Parties at both the far left and far right post more about migration than centrist actors.
D. Migration is mentioned more often in countries with low immigration and rarely in high-immigration countries.

Mock questions week 3:

3.1 Inductive vs. deductive

A core advantage of supervised machine learning over dictionary approaches is…

A. It guarantees perfect accuracy.
B. It requires no validation.
C. It learns patterns from labeled examples automatically
D. It eliminates preprocessing needs.

3.2 Overfitting

Overfitting occurs when…

A. A model fits training data extremely well but generalizes poorly.
B. A model cannot learn patterns in training data.
C. Hyperparameters are randomly selected.
D. Training data are perfectly representative.

3.3 Validation

Why is it important to use a separate test set when validating a machine-learning classifier?

A. Because the test set automatically improves the model's accuracy during training.
B. Because the model cannot learn patterns without a separate test set.
C. Because evaluating on unseen data estimates how well the model generalizes beyond the training data.
D. Because using the same data for training and testing causes overfitting.

3.4 Hyperparameters

Hyperparameters such as number of hidden layers or learning speed require tuning because…

A. Their optimal values are unknown and task-dependent.
B. They are learned automatically during training.
C. They do not influence performance.
D. There are good theoretical reasons to choose one value over another.

3.5 Embeddings

Word embeddings rely on the idea that…

A. Words are represented as high-dimensional sparse vectors.
B. Words appearing in similar contexts have similar meanings.
C. Word order is irrelevant.
D. Words can be manually clustered by experts.

3.6 Neural text classification

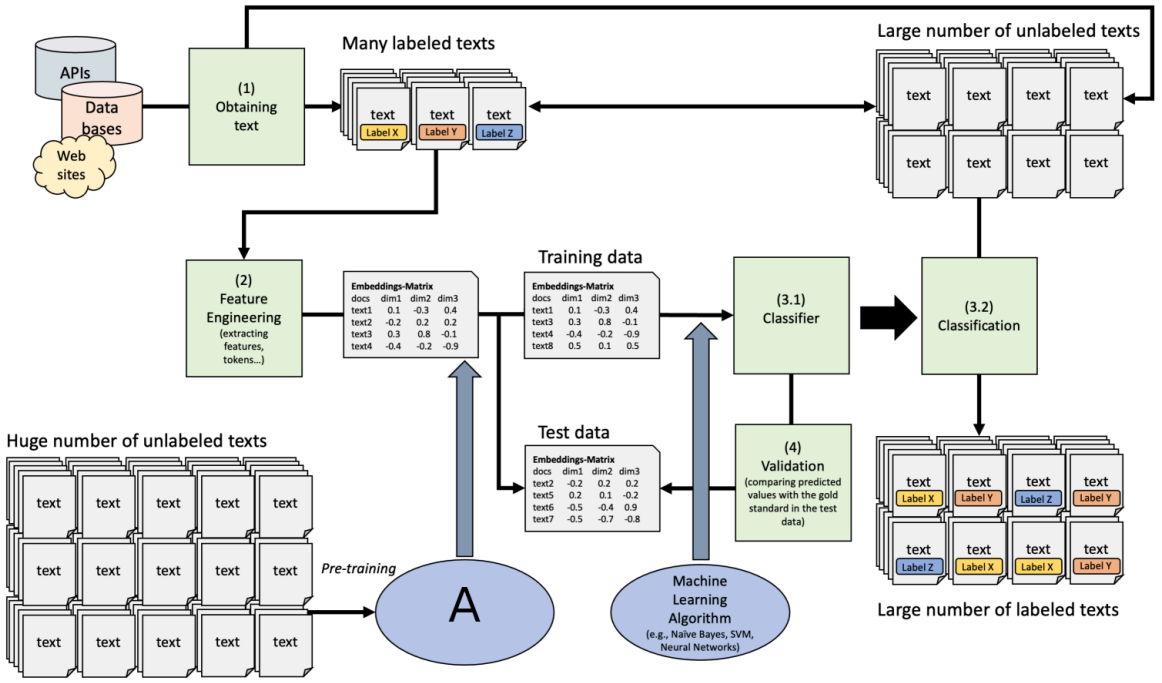
A key advantage of embeddings over bag-of-words in neural networks is…

A. They make the input more sparse..
B. They reduce environmental cost.
C. They eliminate the need for tokenization.
D. They capture semantic similarity.

Mock questions week 4:

4.1 The text classification pipeline

The slides included the following figure, with the contents of the first blue ellipse replaced by "A"



What kind of classification does this describe, and what should be in the place of "A"?

A. Text classification with transformers, encoder only. A are Word-Embeddings, e.g. from BERT
B. Text classification with transformers, encoder only. A is a Large Language Model, e.g. GPT
C. Text classification with transformers, decoder only. A are Word-Embeddings, e.g. from BERT
D. Text classification with transformers, decoder only. A is a Large Language Model, e.g. GPT

4.2 Word embeddings

What is a key difference between early word embeddings (e.g., Word2Vec, GloVe) and embeddings produced by an encoder like BERT?

A. Early embeddings require supervised training, while BERT is trained without any text data.
B. Early embeddings assign each word a single fixed vector, while BERT generates different embeddings for the same word depending on its context.
C. BERT produces smaller embedding vectors than Word2Vec or GloVe.
D. Early embeddings capture syntax and long-range dependencies better than BERT.

4.3 Self-attention

What is the main function of the self-attention layer in transformer-based language models?

A. To determine how strongly each word in a sequence relates to every other word
B. To remove irrelevant words from the input sequence.
C. To ensure attention is based on the left-to-right reading order of the sentence.
D. To convert text directly into class labels without additional processing.

4.4. Transformer layers

Multiple transformer layers allow the model to…

A. Capture increasingly complex patterns in text.
B. Memorize training data verbatim.
C. Reduce model size.
D. Avoid backpropagation.

4.5 Zero-shot vs few-shot

Few-shot learning differs from zero-shot learning because it…

A. Requires a pretrained model with more parameters (e.g. GPT-4 instead of GPT-3).
B. No longer needs to use a prompt.
C. Uses contextual embeddings rather than word frequencies.
D. Uses a small set of labeled examples in the prompt.

4.6. Text Analysis methods

Below you see four statements about different computational text-analysis methods.
Match each statement (1–4) with the method it best describes (A–D).

Methods:
A. Dictionary Approaches
B. Supervised Machine Learning
C. Word Embeddings
D. Large Language Models (LLMs)

Statements:

1. _____ performs tasks using massive pretrained models without task-specific training data
2. _____ uses predetermined rules to assign categories;
3. _____ represents words as dense vectors learned from context, capturing semantic similarity.
4. _____ learns patterns from labeled examples to predict categories for new text

Answer key:

Week 1: C, A, D, B

Week 2: B, A, A, A, D, C

Week 3: C, A, C, A, B, D

Week 4: A, B, A, A, D
4.6: 1D, 2A, 3C, 4B