UNIVERSITY OF GOTHENBURG

SPRÅKBANKENTEXT

Harnessing Artificial Intelligence to Combat Disinformation

Ricardo Muñoz Sánchez

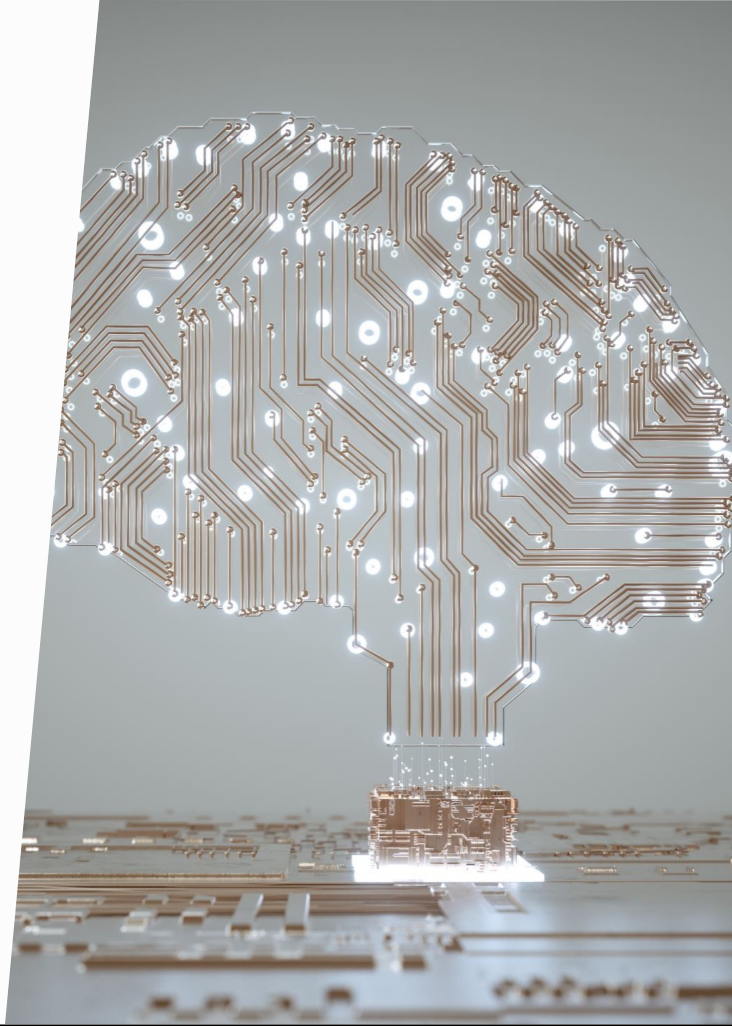
PART 1
A (Very) Brief Introduction to AI

A.K.A. How do computers do language?

Artificial Intelligence

- Artificial intelligence is when a machine can learn from data and generalize from there
  - Not to be confused with general artificial intelligence!

- There are several areas within AI:
  - Computer Vision
  - Natural Language Processing
  - Speech Recognition
  - Machine Learning



What Do We Mean by NLP?

- NLP is an acronym for Natural Language Processing

- Also known as computational linguistics or language technology

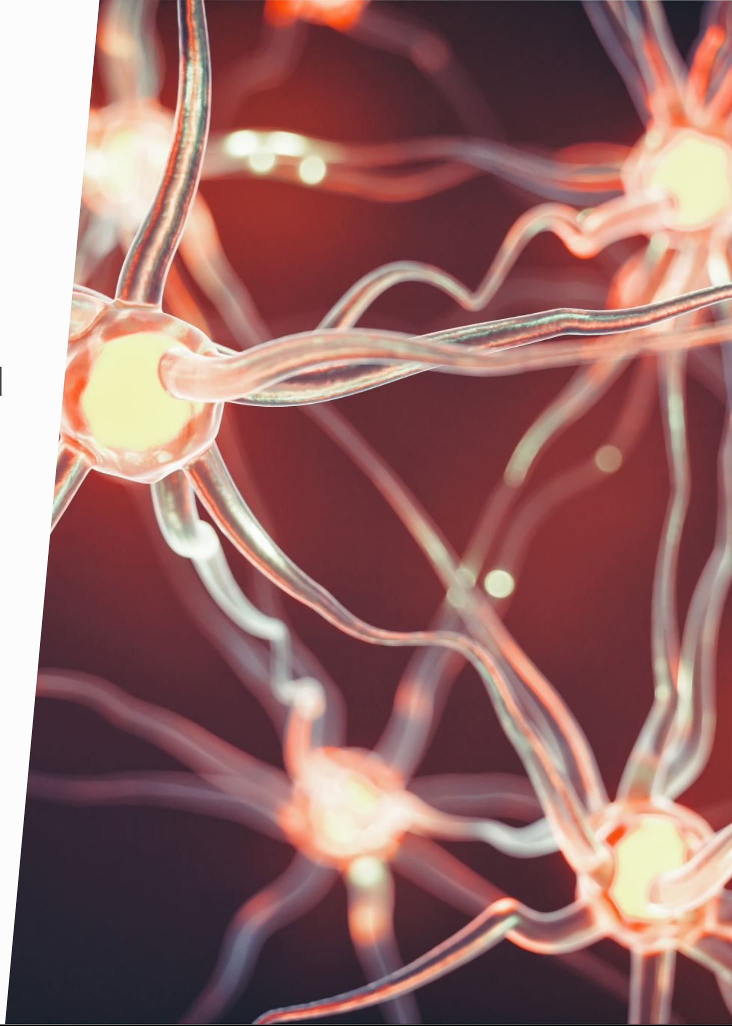
- We use computational methods to study language

What Can We Do with NLP?

Text classification	For example hatespeech detection, sentiment analysis, etc.
Automatic translation	Think of Google Translate and DeepL
Natural language generation	Think of chatbots like Chat-GPT, text summarisation, etc.
Many more!	Natural language understanding, apps like Grammarly, etc.

How Do We Do This?

- Symbolic approaches
  - We design a set of rules and representations to model how language works
- Traditional machine learning
  - We extract some features and use mathematical models to learn and adapt to them
- Deep learning
  - We feed insane amounts of data to a neural network so that it learns its own rules and representations



PART 2
What *is* Disinformation?

And Why Should We Care?

What are “Fake News”?

- The term “fake news” is not well defined!
- It has been used as:
  - A general term for disinformation
  - A term for intentionally false news
  - A way to disqualify journalistic outlets

Misinformation – False information that is spread, regardless of intent.

Disinformation – False information spread with the intent to deceive or to manipulate.

Some Relevant Terms

Rumours

Clickbait

Propaganda

Satirical News

Fake/False News

Biased News

The Problem with Intent

- Most of these definitions hinge on intent

- However, intent is hard (if not impossible) to establish

- This complicates gathering data in a reliable and consistent manner

Ok, but where do I get my data from?

- Expert annotators
  - Fact-checking organizations for article-level annotations
  - Watchdog organizations for source-level annotations

- Crowdsourcing
  - Asking non-experts to annotate data

Ethical Concerns

- Where do we draw the line between policing and censorship?
- Who is telling us what is true and what is false?
- Can we *really* detect falsehood just through text?

PART 3
Disinformation and AI

How Do We Use AI to Study Mis/Disinformation?

How Do We Use AI to Stop Disinformation?

To identify (intentionally) misleading content

- Fake and/or biased news detection
- Detection of doctored images / deepfakes

To help fact-checkers

- Through automatic fact-checking
- Flagging articles/posts/trends where fake news or other kinds of disinformation might appear

To study how disinformation evolves over time

- Analysing how a specific piece of disinformation changes over time
- Tracking the spread of fake news and rumors, both in social media and through different sites

Three Different Approaches

Knowledge-based

Compare the information in a text against a knowledge base

Content-based

Check for cues of deception in the style of the text

Context-based

Analyse the context in which the article exists (e.g. social media interactions)

Knowledge-Based Approaches

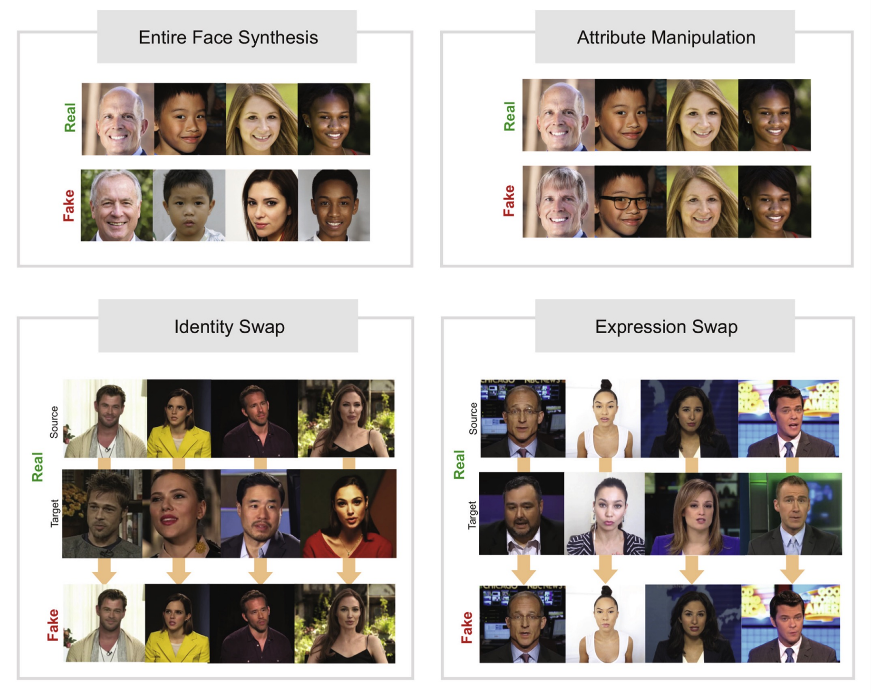
- Automated fact-checking
  - Given a claim, verify its veracity with a knowledge base
  - Identifying previously fact-checked claims

- Possible issues
  - Can be too slow
  - Might not work with breaking news

Content- and Context-Based Approaches

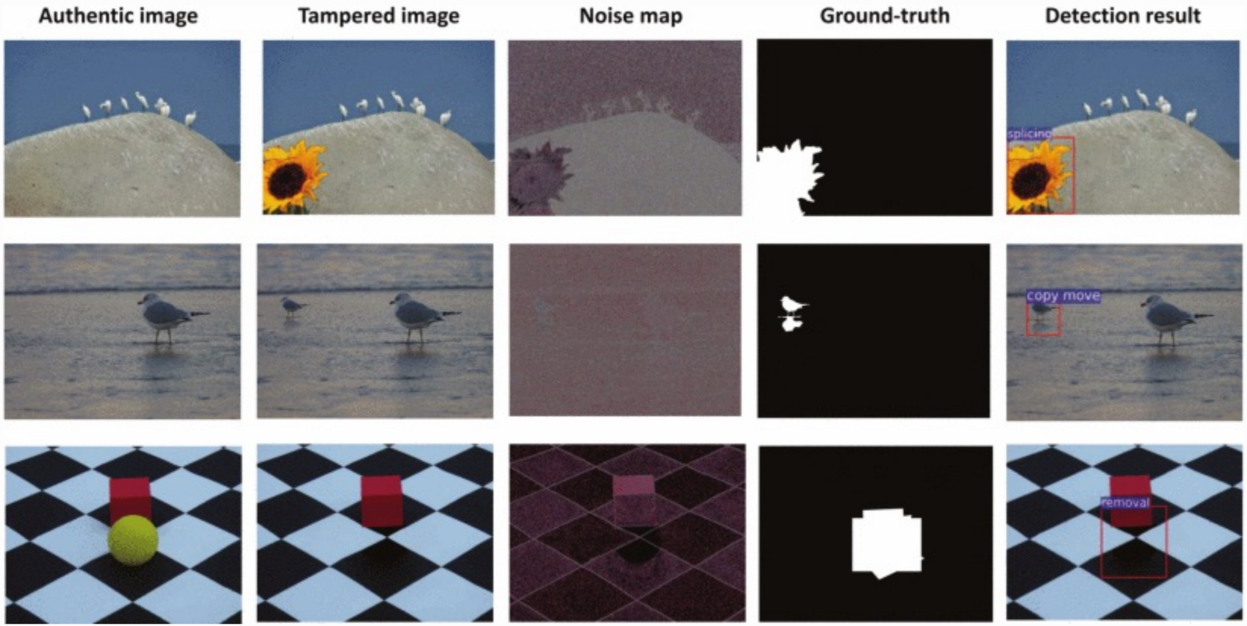
- Are usually focused around machine learning methods
- Use a combination of content- and context-based features
- Can focus on one or more of data, features, and/or models

Deepfakes



From “Deepfakes and beyond: A Survey of face manipulation and fake detection” by Tolosana et al. (2020) [Link]

Detecting Manipulated Images



From “Learning Rich Features for Image Manipulation Detection” by Zhou et al. (2018) [Link]

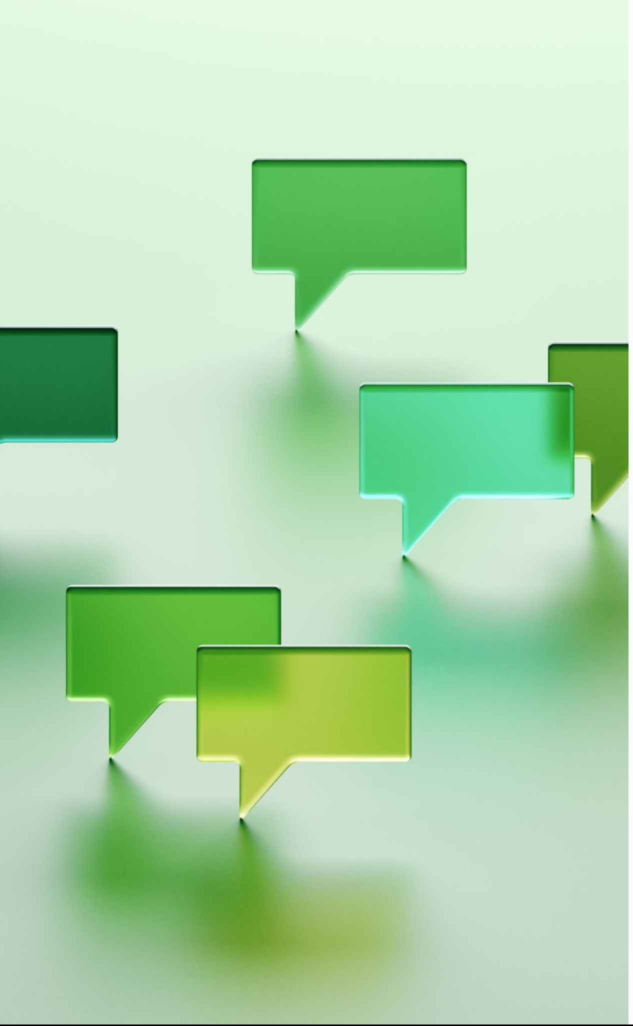
GÖTEBORGS
UNIVERSITET

SPRÅKBANKENTEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@svenska.gu.se
rimusa.github.io

Content-Based Features

- Textual representations
  - TF-IDF
  - Word embeddings
- Linguistic features
  - Distribution of POS, punctuation, etc.
  - Syntactic trees
- Psycholinguistic features
  - Sentiment and emotion analysis
  - Detecting morality and principles, among others



Context-Based Features

- Can be related to the publication of the article
  - Who wrote and who published the article? When and where was it published?
  - Who are the ad partners of the publishing website?
- Can also be related to social network engagement
  - Who was the original poster?
  - How was the article shared/liked/interacted with?
  - Who interacted with the post?

Sources

- Definitions of fake news
  - Allcott, Hunt, and Matthew Gentzkow. “Social Media and Fake News in the 2016 Election.” *Journal of Economic Perspectives* 31, no. 2 (May 2017): 211–36. https://doi.org/10.1257/jep.31.2.211.
  - Tandoc, Edson C., Zheng Wei Lim, and Richard Ling. “Defining ‘Fake News.’” *Digital Journalism* 6, no. 2 (February 7, 2018): 137–53. https://doi.org/10.1080/21670811.2017.1360143.
  - Guess, Andrew M, and Benjamin A Lyons. “Misinformation, Disinformation, and Online Propaganda.” In *Social Media and Democracy: The State of the Field, Prospects for Reform*, Vol. 10. SSRC Anxieties of Democracy. Cambridge, United Kingdom ; New York, NY: Cambridge University Press, 2020.
  - Egelhofer, Jana Laura, and Sophie Lecheler. “Fake News as a Two-Dimensional Phenomenon: A Framework and Research Agenda.” *Annals of the International Communication Association* 43, no. 2 (April 3, 2019): 97–116. https://doi.org/10.1080/23808985.2019.1602782.

Sources

- Definitions of other kinds of mis- and disinformation
  - Pendleton, Susan Coppess. “Rumor Research Revisited and Expanded.” *Language & Communication* 18, no. 1 (January 1, 1998): 69–86. https://doi.org/10.1016/S0271-5309(97)00024-4.
  - Escher, Anna, and Anthony Ha. “WTF Is Clickbait?” *TechCrunch* (blog), September 25, 2016. https://social.techcrunch.com/2016/09/25/wtf-is-clickbait/.
  - Garrett, R. Kelly, Robert Bond, and Shannon Poulsen. “Too Many People Think Satirical News Is Real.” *The Conversation* (blog), August 16, 2019. http://theconversation.com/too-many-people-think-satirical-news-is-real-121666.
  - Bednar, Peter, and Christine Welch. “Bias, Misinformation and the Paradox of Neutrality.” *Informing Science: The International Journal of An Emerging Transdiscipline* 11 (2008): 85–106.
  - Egelhofer, Jana Laura, and Sophie Lecheler. “Fake News as a Two-Dimensional Phenomenon: A Framework and Research Agenda.” *Annals of the International Communication Association* 43, no. 2 (April 3, 2019): 97–116. https://doi.org/10.1080/23808985.2019.1602782.

Sources

- Various literature reviews
  - Bondielli, Alessandro, and Francesco Marcelloni. “A Survey on Fake News and Rumour Detection Techniques.” *Information Sciences* 497 (September 1, 2019): 38–55. https://doi.org/10.1016/j.ins.2019.05.035.
  - Oshikawa, Ray, Jing Qian, and William Yang Wang. “A Survey on Natural Language Processing for Fake News Detection.” In *Proceedings of the 12th Language Resources and Evaluation Conference*, 6086–93. Marseille, France: European Language Resources Association, 2020. https://www.aclweb.org/anthology/2020.lrec-1.747.
  - Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. “Fake News Detection on Social Media: A Data Mining Perspective.” *ACM SIGKDD Explorations Newsletter* 19, no. 1 (September 1, 2017): 22–36. https://doi.org/10.1145/3137597.3137600.

Sources

- Deep Fakes
  - Tolosana, Ruben, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. “Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection.” *Information Fusion* 64 (December 2020): 131–48. https://doi.org/10.1016/j.inffus.2020.06.014.
  - Zhou, Peng, Xintong Han, Vlad I. Morariu, and Larry S. Davis. “Learning Rich Features for Image Manipulation Detection.” In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1053–61. Salt Lake City, UT, USA: IEEE, 2018. https://doi.org/10.1109/CVPR.2018.00116.