DIGITAL METHODS IN HUMANITIES

Zhiru Sun
Associate Professor
University of Southern Denmark

SDU

If you have textual data…

Text Similarity
→ Identify repeated or borrowed text across sources

Topic Modeling and Sentiment Analysis
→ Classify content by topic and sentiment

Information Retrieval
→ Identify key information from texts

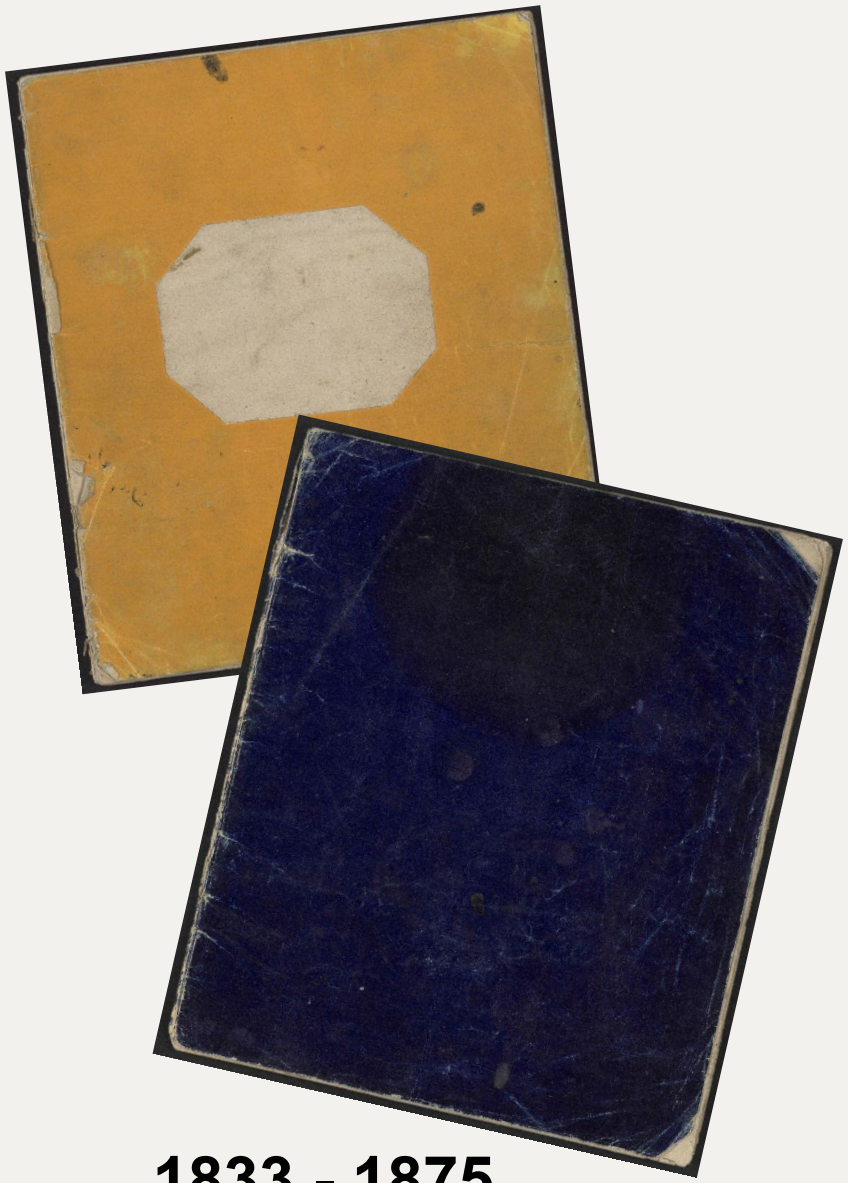
You can turn words into insights with digital tools!

SDU

DETECT TEXT REUSE IN H.C.ANDERSEN'S WORK

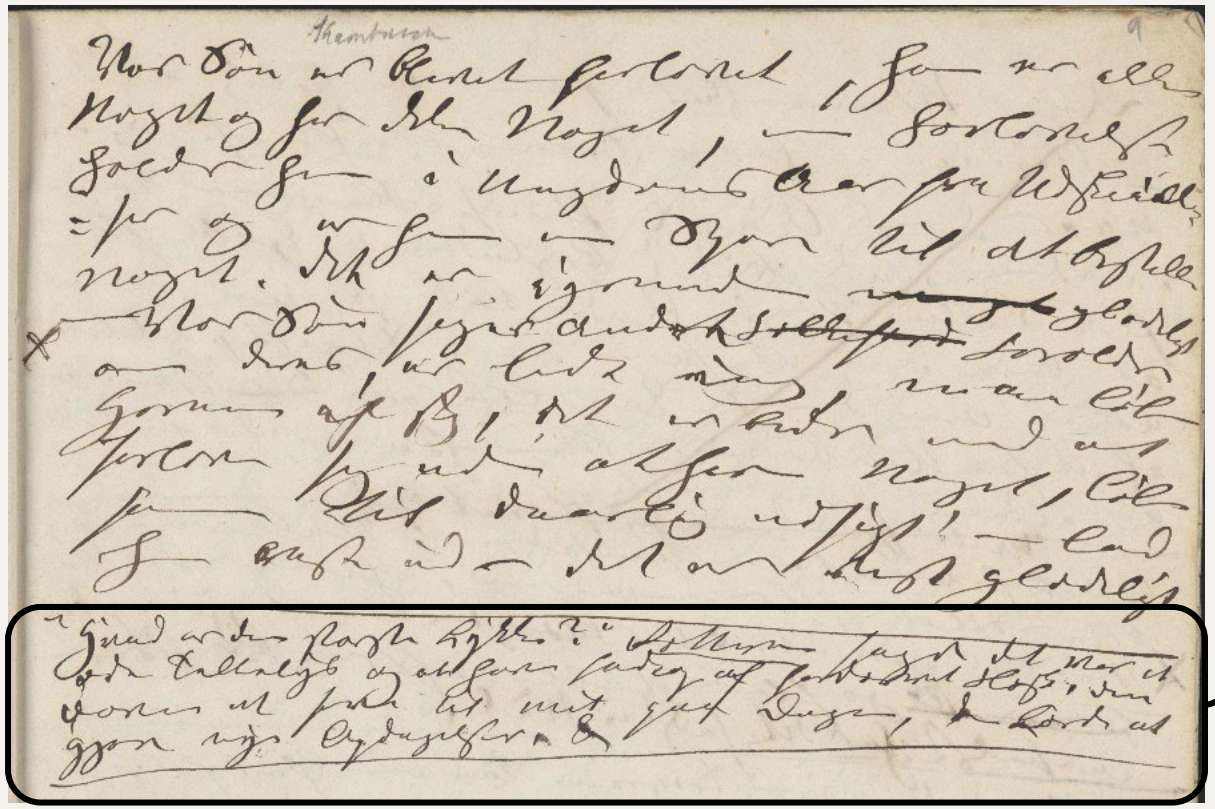
ZHIRU SUN
NILS HOLGER BERG
TARIQ YOUSEF

It-vest
samarbejdende universiteter



1833 - 1875

SDU



Digitalized Note

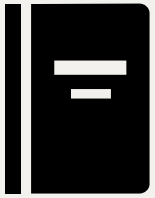
”Hvad er den største Lykke?” Rotterne sagde det var at æde Tellelys og at have fuldtop af fordærvet Flæsk, den Dovne at sove til mit paa Dagen, den Lærde at gjøre nye Opdagelser, &

“Isjomfruen/The Ice Maiden”

"Jeg hørte i går aftes," sagde køkkenkatten, "rotterne siger, at den største lykke er at fortære telleys og fylde sig selv med fordærvet flæsk. Hvem skal man tro, rotter eller kærester?' ‘Heller ikke!’ svarede Stuekatten. 'Det er altid det sikreste!'

SDU

Project Goal



1,017

Notes (1833-1875)

VS

168

Fairytales and Works

Detect text similarities between H.C.Andersen's digitalized notes and published works.

Text Cleaning

Aarets Historie (p_00002)

Mod Aften var det blikstille, Himlen saae ud,
som om den var feiet og gjort mere høi og
gjennemsigtig, Stjernerne syntes splinternye, og
nogle vare saa blaa og klare, – og det frøs saa det
knagede efter, – sagtens kunde da det øverste
Sneelag blive saa stærkt, at det i Morgenstunden
bar Graaspurvene; de hoppede om snart oppe
snart nede, hvor der var skovlet, men megen Æde
var der ikke at finde, og de frøs ordentligt.

Processed Aarets Historie (p_00002)

mod aften være den blikstill himlen saae ud som
om den være fei og gøre meget høi og
gjennemsigtig stjernerne synes splinternye og
nogen vare saa blaa og klar og det frøs saa det
knage efter sagtens kunde da den øvre sneelag
blive saa stærk at det i morgenstunden bar
graaspurvene de hoppe om snart oppe snart nede
hvor der være skovle men meget æde være der
ikke at finde og de frøs ordentligt

Chunking

Processed Aarets Historie (p_00002)

mod aften være den blikstill himlen saae ud som om den være fei og gøre meget høi og gjennemsigtig stjernerne synes splinternye og nogen vare saa blaa og klar og det frøs saa det knage efter sagtens kunde da den øvre sneelag blive saa stærk at det i morgenstunden bar graaspurvene de hoppe om snart oppe snart nede hvor der være skovle men meget æde være der ikke at finde og de frøs ordentligt

p_00002_1 mod aften være det blikstille himlen saae ud som om den være fei og gøre meget høi og gjennemsigtig stjerne synes splinterny og nogen vare saa blaa og klar og

p_00002_2 synes splinterny og nogen vare saa blaa og klar og det frøs saa det knage efter sagtens kunde da den øver sneelag blive saa stærk at det i morgenstund bar

p_00002_3 øver sneelag blive saa stærk at det i morgenstund bar graaspurv de hoppe om snart oppe snart nede hvor der være skovle men meget æde være der ikke at finde

p_00002_4 være skovle men meget æde være der ikke at finde og de frøs ordentligt

SDU

Text Processing

1,017
Notes

5,355
Chunks

VS

251,920,655
Pairwise Comparison

171
Works

47,061
Chunks

SDU

Similarity Calculation: Lexical Similarity

→**Lexical similarity** is a measure of the degree to which the word sets of two given texts are similar.

→A lexical similarity of **1** (or 100%) indicates a total overlap between vocabularies, whereas **0** means there are no common words.

→N-grams overlap

→Unigram, Bigrams, Trigrams, 4-grams, 5-grams, etc.

SDU

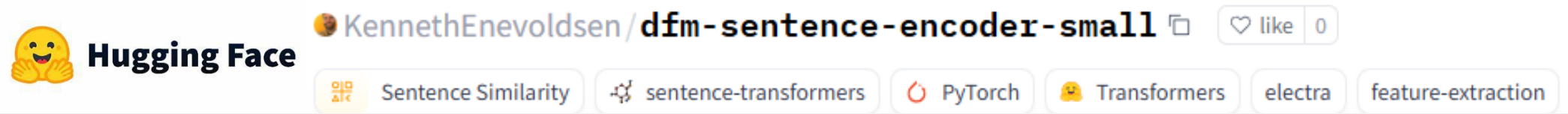
Similarity Calculation: Semantic Similarity

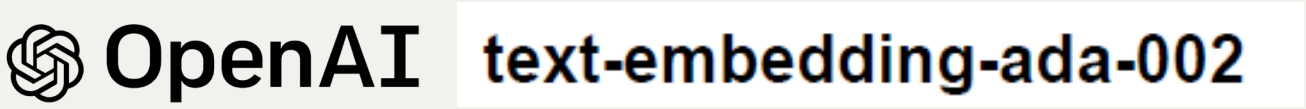
→**Semantic Similarity** determines the similarity of a given text from a combination of semantic and syntactic information.

→Large Language Models (LLMs)

→***Hugging face***: dfm-sentence-encoder-small

→***OpenAI***: text-embedding-ada-002





SDU

HCA Text Reuse Feedback Collection Tool Home Stats

Similarity Scores

Lexical Overlap: Tokens 30.00% | Bigrams 0.00% | Trigrams 0.00% | 4Grams 0.00% | 5Grams 0.00% | Named Entities 0

Longest Common Substring: Original 2 | Lemmatized 1 | Lemmatized without Stopwords 1

Semantic Similarity: Original 74.07% | Lemmatized 58.96% | OpenAI 89.08%

41-4_II-6_div627 (Ingen brug udpeget / no uses known yet)

jeg var i familien men jeg groede ikke fast vi samtalede som i en omnibus kjendte hinanden som i en omnibus generede hinanden ønskede at den naboe snart var afsted

Jeg var i Familien men jeg groede ikke fast, vi samtalede, som i en Omnibus, kjendte hinanden, som i en Omnibus, generede hinanden, ønskede at den Naboe snart var afsted; der sad Tante, med smalle Læber, Næsen opad, som saae den stolt over hele Personen, og sagde vi er Herskabet. dvask, Sønnerne flaue, man passer ikke til hinanden, glade ved efter Bordet at falde fra hinanden, almen er som Lampen der lyser op i Huset, er hen er det mørkt. –

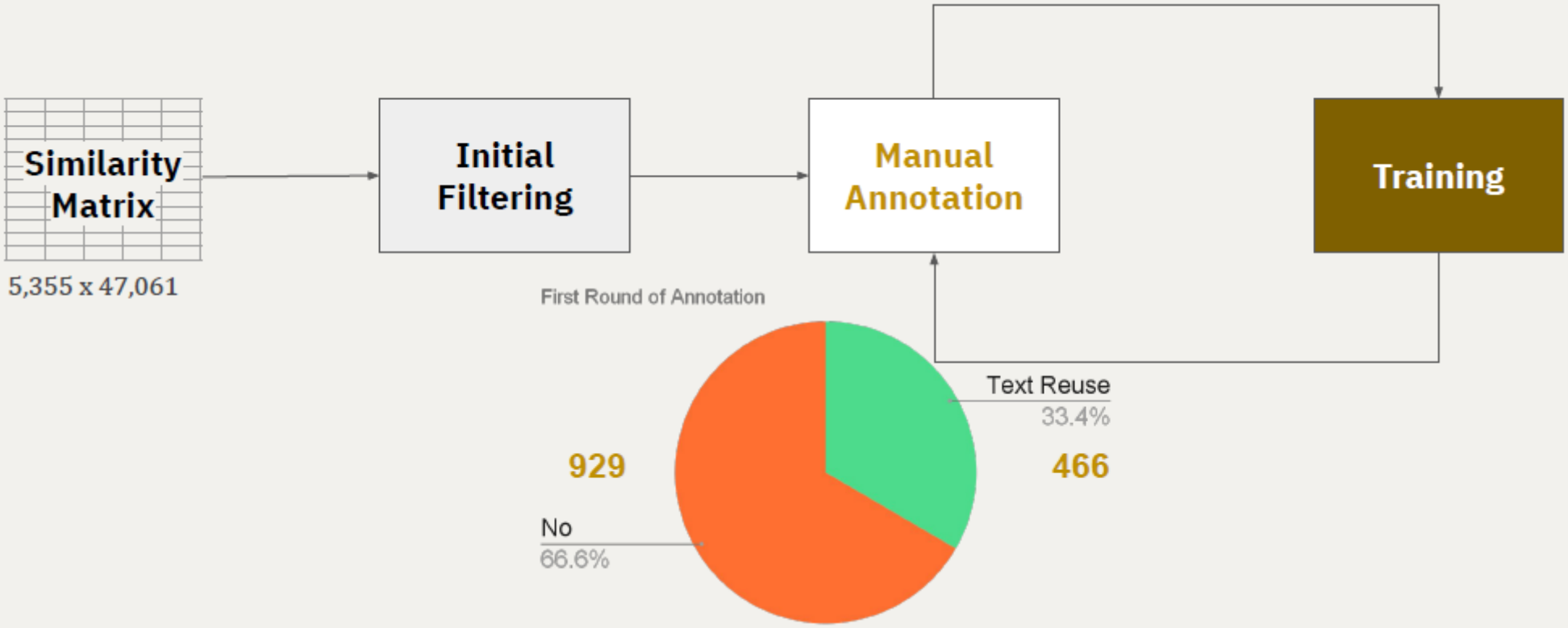
Pebersvendens Nathue__p_00028

anden men vi kunne forstaae det man kunne være i hus i familie og gro dog ikke fast man samtale som man samtale i en postvogn kjende hinanden som man

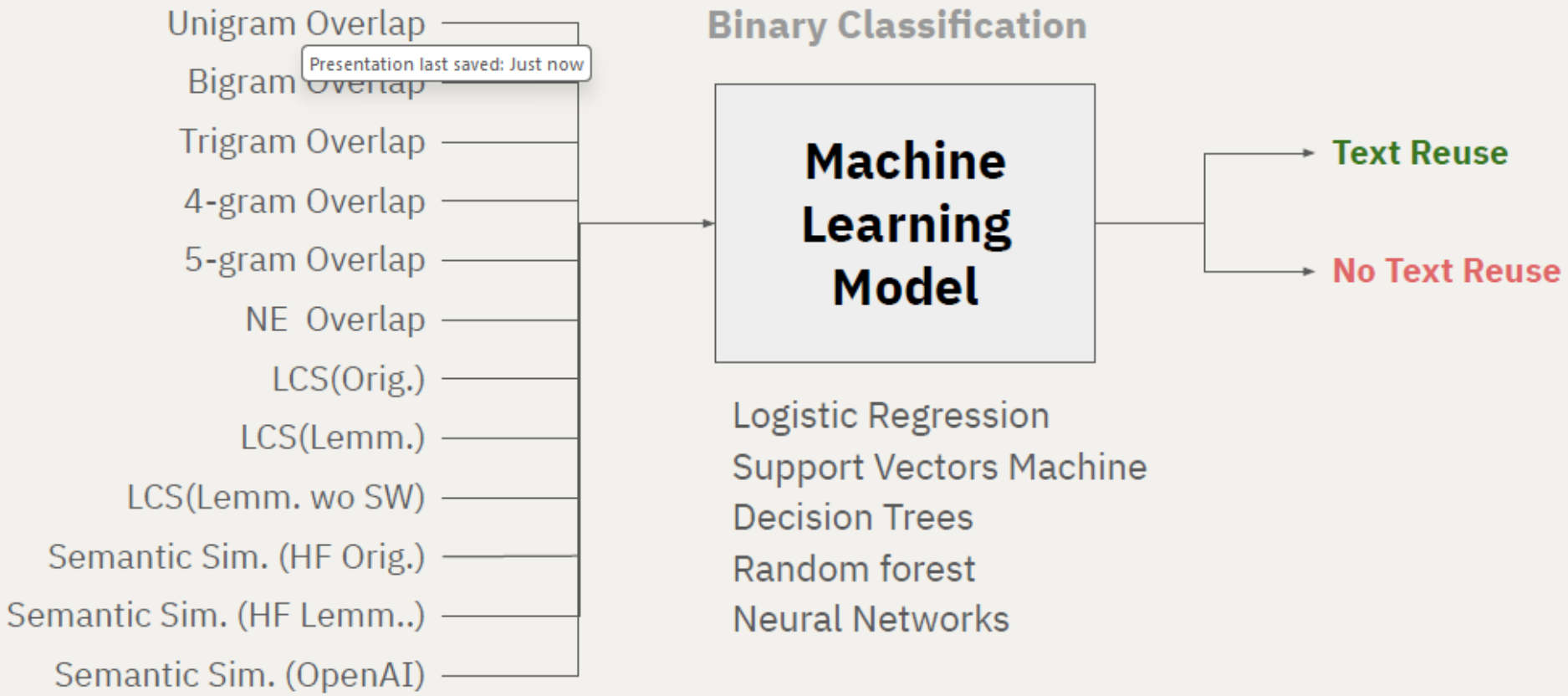
Velkommen fik han, Viin fuldt op i Bægeret, muntert Selskab, fornemt Selskab, en hyggelig Stue og en god Seng, og dog var der slet ikke som han havde tænkt og drømt sig! han forstod ikke sig, han forstod ikke de Andre; men vi kunne forstaae det! Man kan være i Huset, i Familie, og groer dog ikke fast, man samtaler, som man samtaler i en Postvogn, kjender hinanden, som man kjender en Postvogn, generer hinanden, ønsker at man var afsted, eller at vor gode Nabo var afsted. Ja saadant Noget fornam Anthon .

SDU

Manual Annotation



ML Model Building



SDU

Model Evaluation

	Not Text Reuse			Text Reuse			Accuracy
	Pre	Rec	F1	Pre	Rec	F1	
Random Forest	87%	93%	90%	84%	73%	78%	**86%**
Support Vector Machine	67%	100%	80%	0%	0%	0%	67%
Logistic Regression	80%	92%	86%	77%	55%	64%	79%
Decision Tree	85%	84%	84%	68%	71%	69%	79%
Neural Network	0%	0%	0%	33%	100%	50%	33%

Results I

41-4_II-4_div509	**95**
De røde Skoe	1
De vilde Svaner	1
Fem fra en Ærtebælg	1
Hørren	1
Hyrdinden og Skorsteensfeieren	1
Improvisatoren	1
Sølvskillingen	89
41-4_II-4_div516	**46**
De vilde Svaner	1
Dryaden	1
Iisjomfruen	37
Improvisatoren	1
Kun en Spillemand	1
Lille Claus og store Claus	1
Lykke-Peer	1
Mit Livs Eventyr (inklusive Fortsættelse og Tillæg). 1855-1877	1
Ole Lukøie	1
Venskabs-Pagten	1
41-4-VI-I-31-B_div199	**38**
»At være eller ikke være«	38
41-4_VI-48_div228	**22**
Den lille Havfrue	1
Folkesangens Fugl	20
Iisjomfruen	1

41-4_II-4_div509

Hvad Byen fortæller. Hvad Skoven fortæller veed hvert kunstnerisk Gemyt, der har dygget sig ned under Træernes grønne Løv, og baaret af Anemoner og Skovmærker, men hvad Byen fortæller ja, det veed Alle og Enhver, hvo som haver Ører han hører, og her er at høre og see hvad Alle kunne forstaae, og dog forstaae de ikke all hvad Byen fortæller. som det See nu i ud paa Gaden, der ligge nogle Drenge og grave i Rendestenen, de vil gribe deres Lykke, som den Voxne vil det og een griber ned og finder stikker sig paa en sort Knappenaal, Een skjærer sig paa et Glasskaar, men den tredie er et Lykkens Barn, han griber ned og finder en Sølvskilling og viser med Jubel sit rige Fund, saa bliver han dænget til af de Andre, det er et Billed i det Smaa af hvad der sker i Verden i det større.

SDU

Results II

Work Title	Numb
»At være eller ikke være«	222
Sølvskillingen	89
Mit Livs Eventyr (inklusive Fortsættelse og Tillæg). 1855-1877	70
Iisjomfruen	61
Folkesangens Fugl	30
De to Baronesser	21
Aarets Historie	19
Et godt Humeur	14
Det nye Aarhundredes Musa	14
O. T.	13
Sneglen og Rosenhækken	11
Veirmøllen	11
Kun en Spillemand	10
Tante Tandpine	9
Improvisatoren	9
Laserne	9
Portnøglen	9
Lykke-Peer	8
Vinden fortæller om Valdemar Daae og hans Døtre	8
Skarnbassen	6
Flyttedagen	6
Gudfaders Billedbog	6

Similarity Scores

Lexical Overlap: Tokens 42.86% | Bigrams 30.77% | Trigrams 16.67% | 4Grams 9.09% | 5Grams 0.00% | Named Entities 0

Longest Common Substring: Original 5 | Lemmatized 4 | Lemmatized without Stopwords 4

Semantic Similarity: Original 72.92% | Lemmatized 66.67% | OpenAI 88.37%

41-4_II-5_div536 (Ingen brug udpeget / no uses known yet)

derinde signalerne der lød ud og mældte at fjender kom den hele natlige spændende scene jeg hørte om den gamle bedstemoder staa der med sine børnebørn stode paa veien da

Jeg kom slet ikke derover, uagtet jeg alle tre Aarene var i Fyen paa Glorup og daglig traf paa Folk, som kom derfra og drog dertil, Nysgjærrige og Slægtninge, som saae til deres Kjære. Men som en Duft, kom det Skjønne fra Krigs Skuepladsen og gik op i min Tanke, en ung Læge fortalte mig om sin Marsch med Soldater over de nøgne, lange Gader, hvorledes han et Sted fik en Kirke til sit Laceret, hvor halv mørkt der var derinde; Signalerne der lød ud og mældte at Fjender kom, den hele natlige spændende Scene. Jeg hørte om den gamle Bedstemoder staa, der med sine Børnebørn stode paa Veien da vore Tropper kom og havde strøet Sand og Blomster og raabte med de Smaa: Gud velsigne de Danske; jeg hørte om det Naturspil at der i en Bondes Have voxte røde Valmuer med hvidt Kors, fuldkommeligt Dannebrogen. Een af mine Venner var paa Als og tog da over til Dyppel, hvor alle Huse stode med Revner og Huller ved Kanonskud og Kardetscher og dog stod endnu paa eet af Husene, Fredens Symbol: en Storkerede med hele Familien; den stærke Skyden, Ild og Røg havde ikke jaget Forældrene bort fra Ungerne, da de endnu ikke kunde flyve.

Mit Livs Eventyr (inklusive Fortsættelse og Tillæg). 1855-1877_p_00202

og gå op i min tanke jeg høre om en gammel bedstemoder der med sin børnebørn stode paa veien da vores troppe komme hun have strøe sand og blomste for

Største Delen af Sommeren tilbragte jeg paa Glorup , var der i Vaar og Høst, og blev saaledes Vidne til Svenskernes Ankomst og senere Bortreise. Selv kom jeg ikke til Krigsskuepladsen, jeg blev paa Glorup , hvor der daglig indtraf Folk, som kom derovre fra, Nysgjerrige og Slægtninge, som saae til deres Kjære. Men som en Duft kom det Skjønne fra Krigsskuepladsen og gik op i min Tanke; jeg hørte om en gammel Bedstemoder, der med sine Børnebørn stode paa Veien, da vore Tropper kom, hun havde strøet Sand og Blomster for dem og raabte raabte med de Smaa: »Gud velsigne de Danske!« jeg hørte om det Naturspil, at der i Slesvig i en Bondes Have voxte røde Valmuer med hvide Kors, fuldkommeligt Dannebrog. En af mine Venner besøgte Als og tog derfra over til Dyppel , hvor alle Huse stode med Revner og Huller af Kanonkugler og Kartetscher , og dog stod endnu paa eet af Husene Fredens Symbol: en Storkerede med hele dens Familie; den stærke Skyden, Ild og Røg havde ikke jaget Forældrene bort fra Ungerne, da de endnu ikke kunde flyve.

SDU

Takeaways

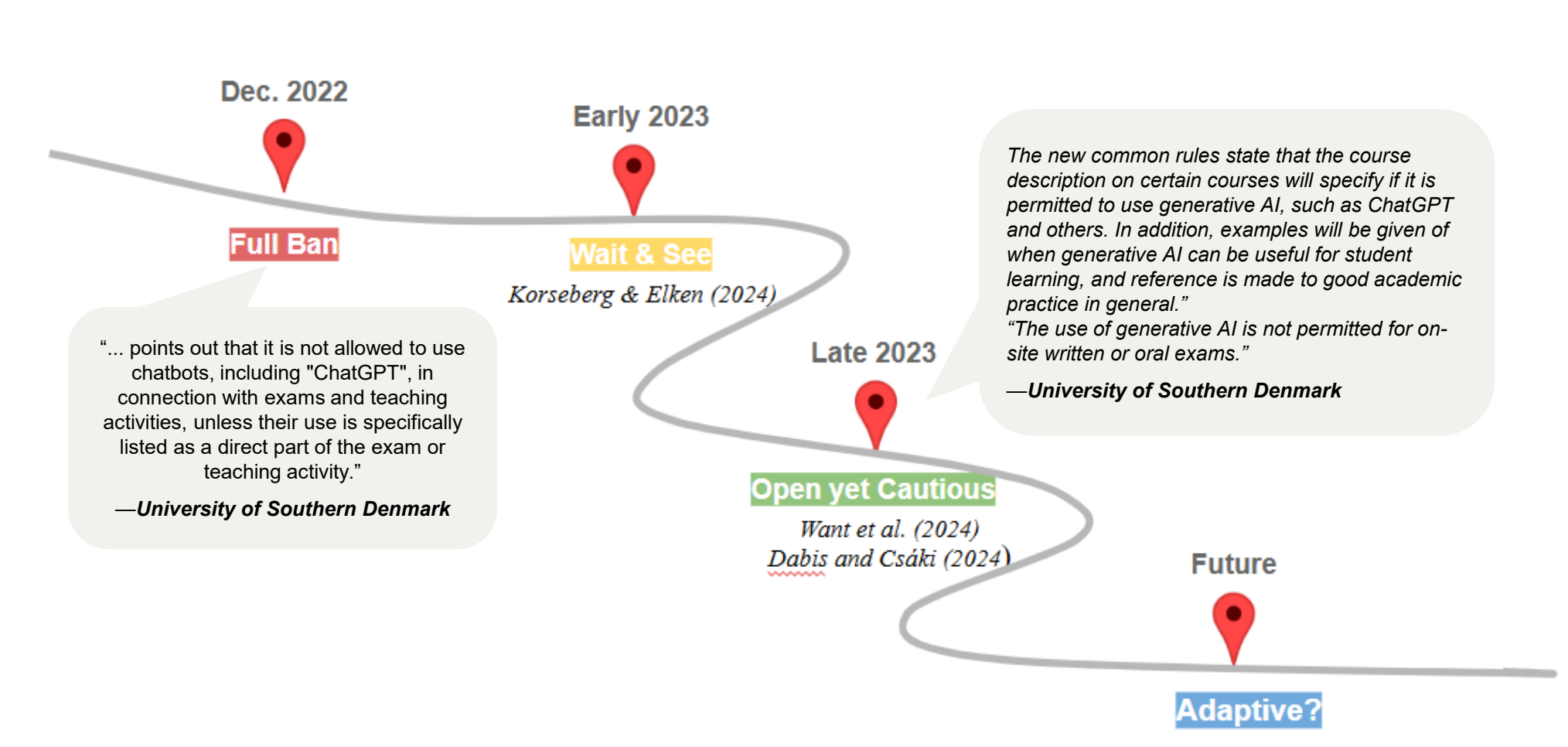
→**Contribution**:
- →Used machine learning and natural language processing techniques to detect text reuse between H.C.Andersen's notes and published works,
- →Help humanities researchers focus on deeper analysis of the text reuse
- →Involved master students in the process and inspired humanities students to engage with digital humanities projects

→**Future work:**
- →Multiple iteration of human annotation to enhance model accuracy
- →Design data visualization tools to demonstrate text reuse

EXPLORATIONI OF GAI POLICY TEXTS IN NORDIC HIGHER EDUCATION INSTITUTIONS

ZHIRU SUN
Nordic EdAI Group

AI Policy Evolution in Higher Education



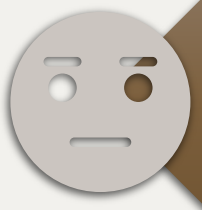
1. Dabis, A., & Csáki, C. (2024). AI and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative AI. *Humanities and Social Sciences Communications*, *11*(1), 1-13.
2. Korseberg, L., & Elken, M. (2024). Waiting for the revolution: how higher education institutions initially responded to ChatGPT. *Higher Education*, 1-16.
3. Wang, H., Dang, A., Wu, Z., & Mac, S. (2024). Generative AI in higher education: Seeing ChatGPT through universities' policies, resources, and guidelines. *Computers and Education: Artificial Intelligence*, *7*, 100326.

Research Landscape of AI Policies

- → **Existing Research Challenges:**
  - → Most studies examined "GAI policy" as a broad term (Wang et al., 2024) or focused narrowly on a single dimension, such as ethics (Dabis & Csáki, 2024).
  - → Insufficient exploration of policy complexity across multiple dimensions.
  - → Minimal exploration of internal policy contradictions.
- → **Our Project:**
  - → GAI policies are complex and multifaceted.
    - → Addressing various dimensions/aspects simultaneously, e.g., teaching, assessment.
    - → Emerging internal contradictions between aspects, e.g., encouraging AI tools for teaching while prohibiting their use in assessments.
- → **Project Goal:**
  - → Employed an ***aspect-based analytical framework*** (Sun, Møller, & Dohn, 2025) to gain a nuanced and comprehensive understanding of policy texts.

1. Sun, Z., Møller, A., & Dohn, N. B. (2025) (*Submitted*). Inherent tensions and resulting challenges for students: A mixed-method exploration of Generative AI Policies in higher education. *British Journal of Educational Technology.*

Research Questions



What are the sentiments of GAI policies towards five aspects: GAI, Student, Ethics, Assessment, and Teaching?

Are there tensions between the five aspects within the policies?

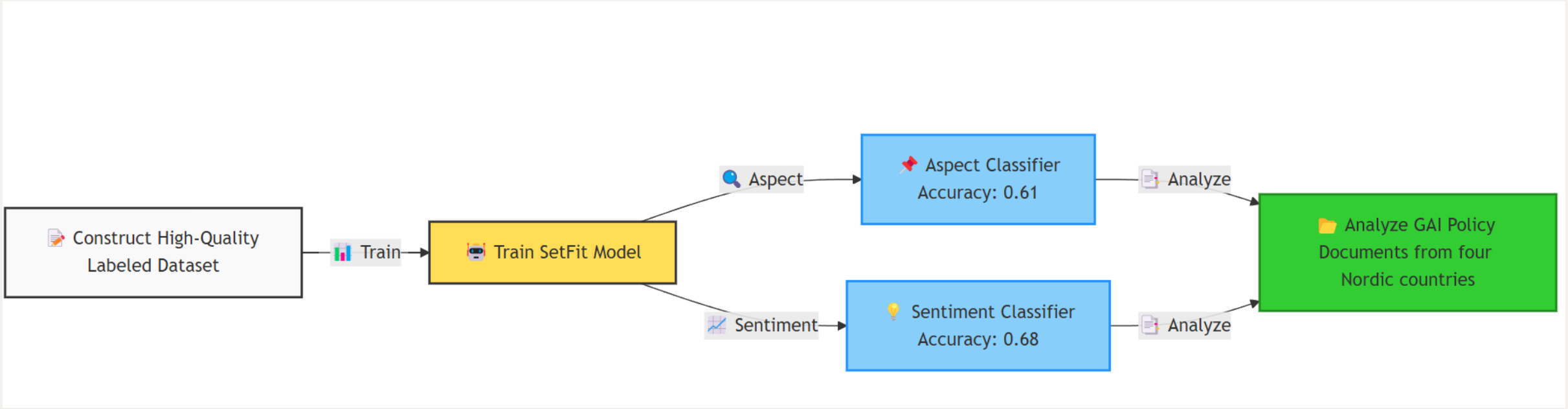
To what level do these senstimentes vary across the four nordic countries: Denmark, Finland, Norway, and Sweden?

SDU

Data and Workflow

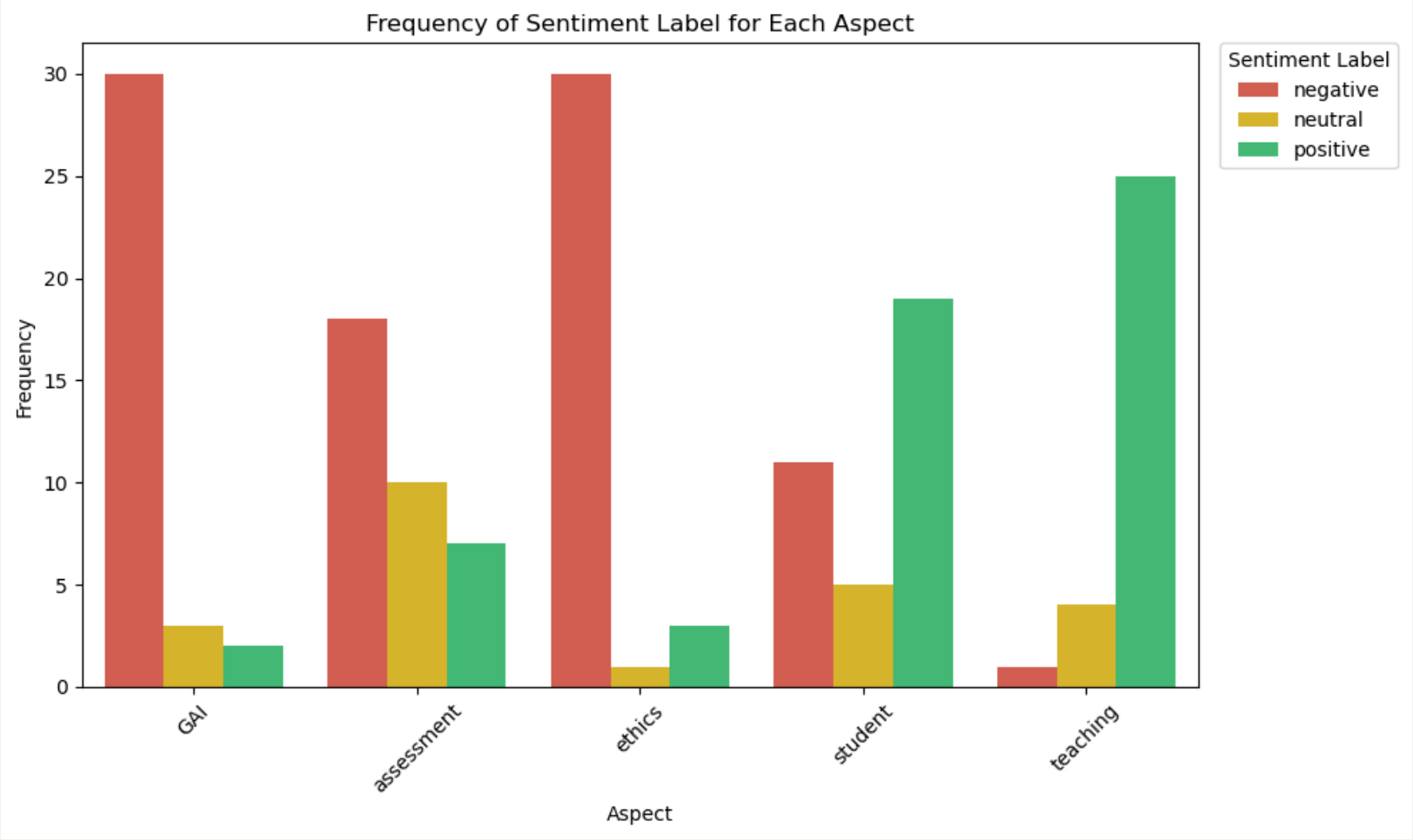
35 policy documents from Denmark (6), Norway (9), Finland (12), and Sweden (8), updated Sep. 2024

84,593 words in the policy documents



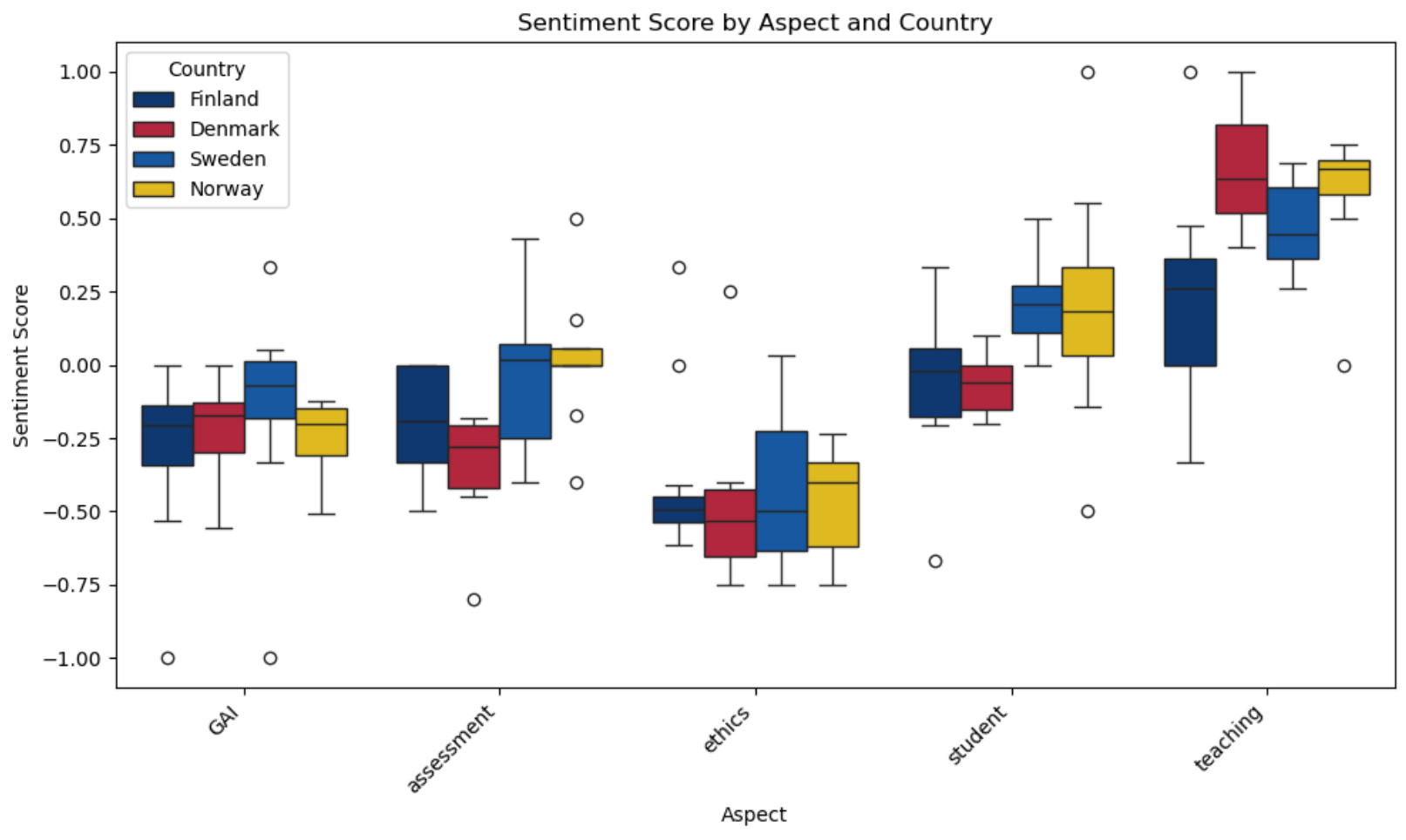
SDU

Main Findings I: Sentiments across aspects



SDU

Main Findings II: Sentiment across countries



SDU

Takeaways

→**Contribution:**

- → Use LLM-powered machine learning techniques to identify variation in sentiments between aspects and across countries.
- → Policies tend to be unspecific and unclear.

→**Future work**

- → Enhance the granularity of data analysis
- → Refine the analytical framework
- → Mixed-method approach to gain qualitative insights from faculty and students on policies, actual needs and their practice

DEMANDING SKILLS: A PIPELINE FOR SKILL EXTRACTION

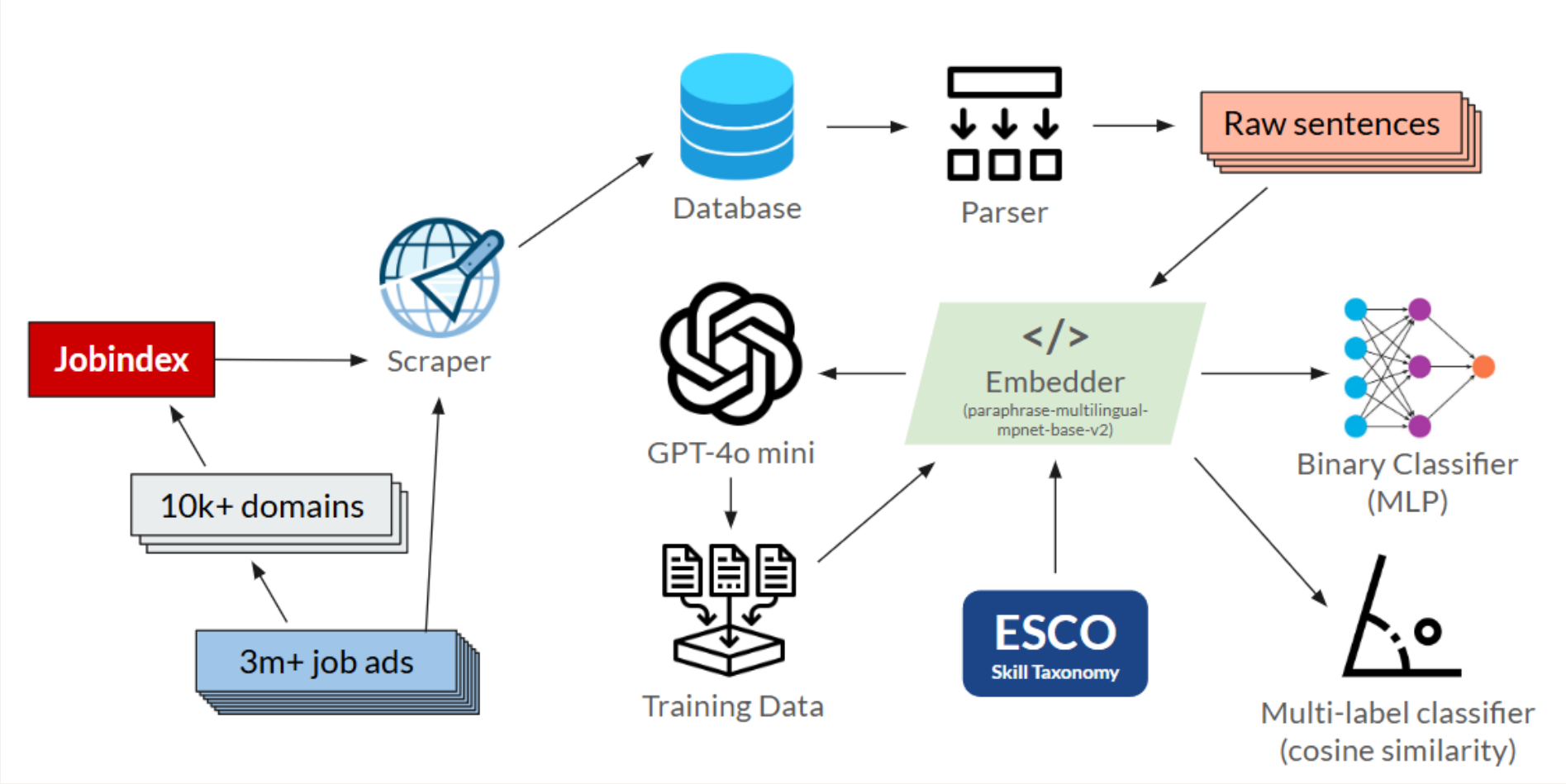
JACOB MØRUP WANG
ZHIRU SUN

Study I: Goal

Develop a systematic approach to identify skills including skill extraction, classification, and representation from Danish job ads (i.e., JobIndex)

Present a case study based on identified skills to understand in-demand competences in the Danish labor market

Systematic Approach: Data Collection



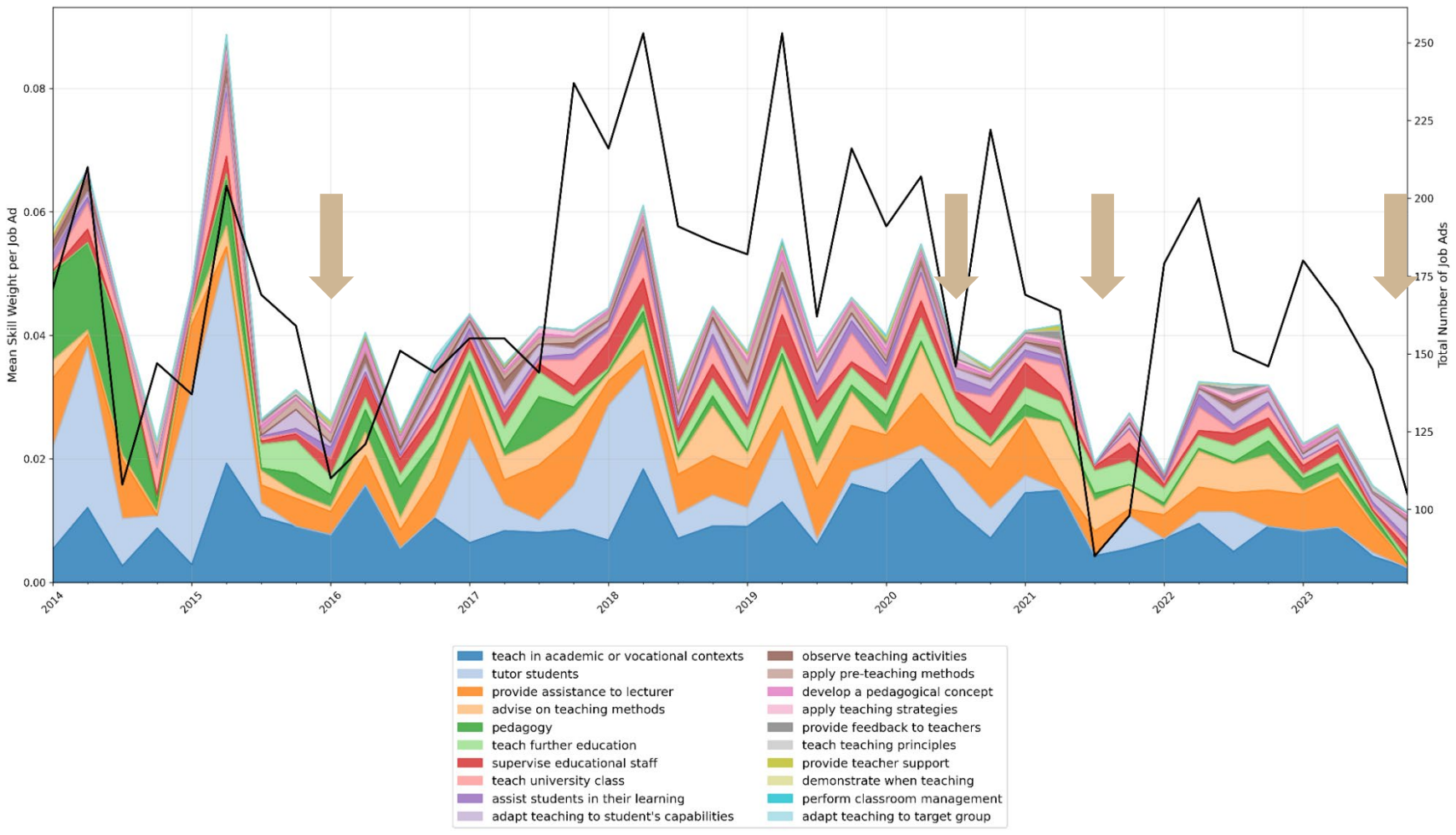
Case Study: SDU

We apply the systematic approach to job ads data from the University of Southern Denmark (SDU) between 2014 to 2023.

Sentences	Skills (score)
Assistant Professor in Aquatic and Wetland Ecology with Emphasis on Ecosystem Services and Nature-based Solutions	aquatic ecology (0.73)
The Elite Centre of Aquatic Nature-based Solutions for climate change adaptation and mitigation based at Department of Biology, University of Southern Denmark in Odense, Denmark, invites applications for a three-year position as assistant professor.	aquatic ecology (0.88)
The starting date is flexible, but we hope to welcome our new colleague on 1 August 2023.	
Description of the position	
We are looking for an assistant professor to join a new team of researchers to be established by the Elite Centre of Aquatic Nature-based Solutions for climate change adaptation and mitigation (Aqua-NbS). Aqua-NbS is a four-year centre funded by the ...	aquatic ecology (0.89)
The position will be affiliated both with Department of Biology , Ecology group , and SDU Climate Cluster.	
At Aqua-NbS we aim to scale up and integrate approaches of Aquatic Nature-based Solutions to expand the provision of multiple ecosystem services, including carbon storage, climate regulation, nutrient removal, coastal protection, and food-web support.	develop aquaculture s (0.78)
Aqua-NbS cover the full land-sea continuum and incorporate a comprehensive range of technological, institutional, and socio-economic settings.	
Aqua-NbS will investigate aquatic and wetland habitats and their valuable ecosystem services to craft inclusive, detailed guidelines on cost-effective opportunities while addressing the challenges in upscaling aquatic NbS.	aquatic ecology (0.69)
About you	
The candidate should have a PhD in biology, ecology, environmental engineering, political science or sociology with an environmental emphasis, environmental/resource economics or similar.	conduct ecological (0.58), environmental ing (0.36)
Experience with mapping of habitats/ecosystem services, nature-based solutions, climate or biodiversity topics, and ecosystem service assessments from any of these disciplinary angles is desirable.	advise on nature cons (0.47)
Expertise in, for example, developing decision support tools, science-policy interface, or stakeholder involvement is also an advantage.	increase the impact of s policy and society (1.00

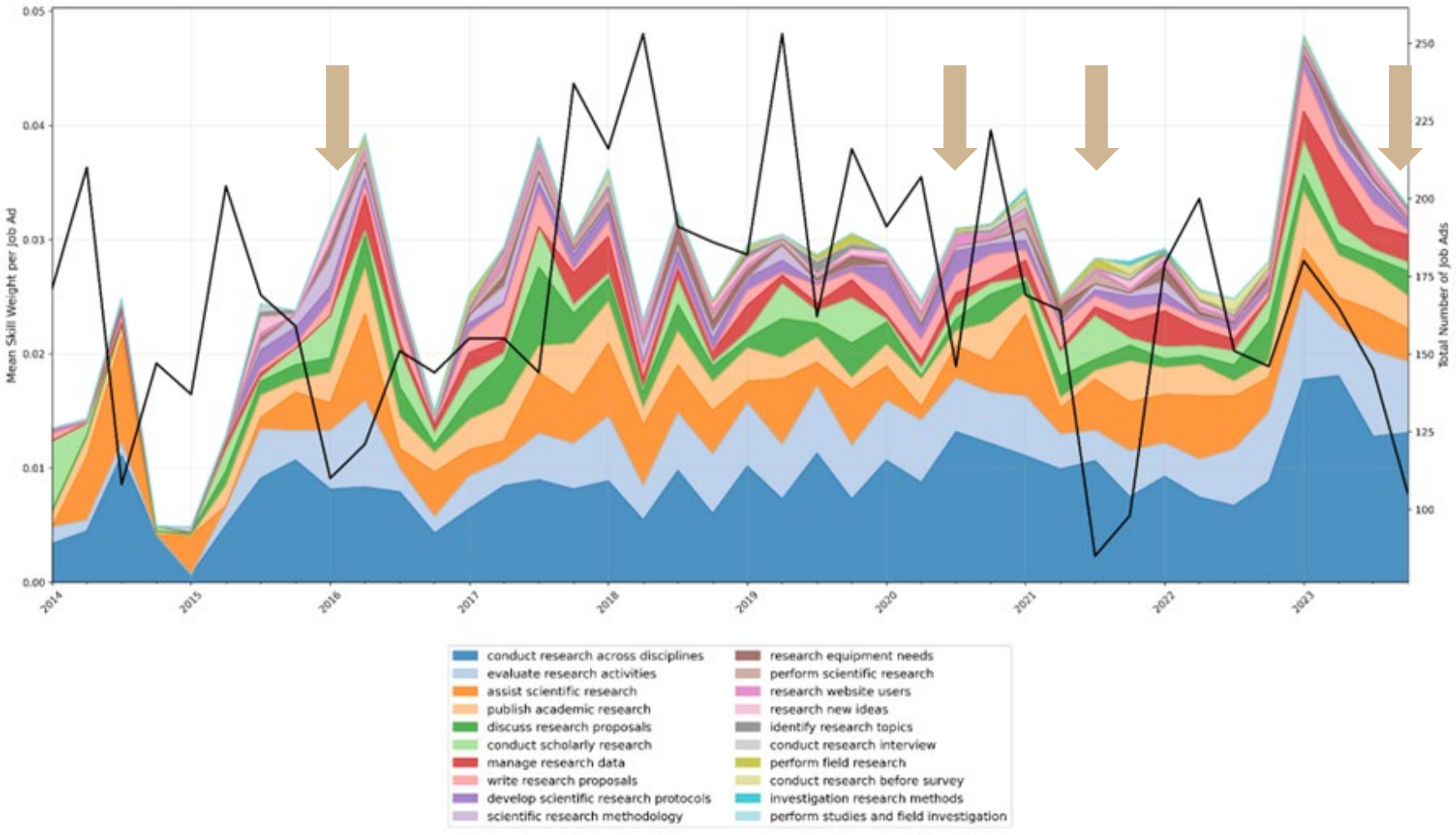
SDU

SDU Case: Skill Demanding Trend in Teaching Roles



SDU

SDU Case: Skill Demand Trend in Research Roles



SDU

Takeaways

→**Contribution**:
- → Our systematic approach successfully collects, cleans, extracts, classifies, and represents skills from job ads.
- → A case study presents a strong potential of our systematic approach.

→**Future work:**
- → Optimizing data cleaning process.
- → LLM-triangulation and/or human-in-the-loop for annotation.
- → Apply the systematic approach to other fields such as healthcare, design, etc.

Thank you!
Questions?

zhiru@sdu.dk

SDU