Blok 2: Natural Language Processing - Competentie Evaluaties

Projectreferenties

Project	Beschrijving	Link
Gemeente Brief Samenvatter	RAG- en Fine-tuning pipeline voor het samenvatten van gemeentelijke brieven	GitLab Repo (backend)

Bewijs van Bijdrage

Bewijs ID	Beschrijving	Type	Locatie/Link	Gerelateerde Competenties
BEW-201	Feedback M&T Opdracht 1b*	Document	feedback-mt-opdracht1b.md	B1, B2, B3, C2
BEW-202	Feedback M&T Opdracht 2a	Document	feedback-mt-opdracht2a.md	B1, B2, B3, C2
BEW-204	Feedback Ontwerp Review Blok 2	Document	feedback-ontwerp-review-b2.md	A1, A2, A3, A4
BEW-205	Feedback Technische Review Blok 2	Document	feedback-tech-review-b2.md	B1, B2, B3, C2
BEW-206	PvA Artikel Blok 2 Feedback	Document	feedback-pva-b2.md	D2, D3
BEW-207	Symposium Presentatie Feedback	Document	feedback-presentation-b2.md	D2, D3

**Note: BEW-201 betreft een deliverable uit blok 1 waarvan de feedback pas in blok 2 beschikbaar kwam en hier is verwerkt.*

A. CONTEXTUALISEREN EN ONTWERPEN

A1: Onderzoek en herformulering van vraagstukken

S/T: Voor het project 'Gemeente Brief Samenvatter' moesten we een oplossing bedenken om complexe juridische teksten in begrijpelijke informatie voor burgers om te zetten. Mijn taak was om de procesflow in kaart te brengen via happy en error flows. **A:** Ik heb flows ontworpen die het hele proces modelleren. De feedback uit de ontwerp-review (BEW-204) was dat de focus te veel op de UI lag en te weinig op de integratie van AI-processen. **R:** De flows zijn nu aangepast om expliciet het onderscheid tussen gebruikersacties en AI-processen te tonen via swimlanes. **T:** In het vervolg ga ik swimlanes gebruiken in mijn flows om expliciet het onderscheid tussen gebruikersacties en AI-processen te tonen.

A2: Requirements opstellen

S/T: Het opstellen van requirements voor een tekst-genererend model bij een gemeente. **A:** Voortbouwend op de feedback uit Blok 1 (BEW-007) wist ik dat bronvermeldingen bij requirements essentieel zijn. Hoewel de uiteindelijke uitwerking in de lijst in dit blok door een groepsgenoot is gedaan omdat ik me op andere technische doelen focuste, heb ik de methodiek (koppeling aan AVG en literatuur) wel bewaakt op basis van de eerdere feedback (BEW-204, BEW-205). **R:** De technische review beoordeelde dit als 'Op verwachting'. De requirements zijn onderbouwd vanuit gebruikersonderzoek, hoewel sommige (zoals TR01) nog concreter en meetbaarder konden (BEW-205). **T:** Om beter individueel deze competentie aan te kunnen duiden is het belangrijk om in het proces van de opstelling van requirements te blijven.

A3: AI-specifieke ontwerprichtlijnen

S/T: Het ontwerpen van mechanismen voor vertrouwen en foutafhandeling bij tekstgeneratie. **A:** Mijn bijdrage bestond uit het maken van de error flows. Op basis van feedback (BEW-204) heb ik de focus verlegd van algemene UI-fouten (zoals een leeg document) naar AI-specifieke fouten, zoals hallucinaties in de samenvatting. **R:** Er is nog werk nodig wat betreft de robuustheid van de error flow in de wiki. De focus moet meer op 'hoe' het systeem herstelt van een foute generatie (BEW-204). Dit is ondertussen verbeterd. **T:** De verbeteringen die zijn gemaakt op basis van de feedback kan ik in de toekomst opnieuw toepassen.

A4: Geschiktheid van AI onderbouwen

S/T: Bepalen waarom AI (en welk type) geschikt is voor het samenvatten van brieven. **A:** We hebben een bewuste keuze gemaakt voor encoder-decoder modellen vanwege het belang van feitelijke correctheid boven creativiteit. Ik heb geholpen met het beargumenteren van deze keuze door het leveren van een centrale paper over de keuze van encoder-decoder modellen. **R:** De review was positief over de mate waarin de keuze van encoder-decoder was verantwoord (BEW-204): "Mooi om te zien dat jullie het belang van de incorrectheid van informatie heel zwaar wegen in jullie beslissing om voor encoder-decoder modellen te kiezen." **T:** Het belang van het analyseren van de literatuur op basis van de benodigde use-case is iets wat ik in de toekomst kan meenemen. Hoewel decoder-only modellen meer populair zijn, heeft de paper laten zien dat encoder-decoder modellen beter zijn voor de gegeven taak.

B. ONTWIKKELEN EN MODELLEREN

B1: Dataset exploratie en preparatie

S/T: Voor zowel de M&T opdrachten als het groepsproject moest data geprepareerd worden. **A:** Cruciaal in dit blok was het verwerken van de feedback uit Blok 1 over **data leakage** (BEW-008). In zowel opdracht 1b als 2a heb ik de dataset eerst gesplitst voordat ik EDA of transformaties uitvoerde (BEW-201, BEW-202). Voor het groepsproject heb ik de EDA en preprocessing uitgevoerd op gemeentelijke PDF-data. **R:** De technische review beoordeelde dit als 'Boven verwachting' (BEW-205). De EDA legde duidelijke knelpunten bloot die de methodiek hebben beïnvloed. De splitsing was methodologisch correct. **T:** De focus voor het eindrapport ligt op het nog transparanter maken van de preprocessing methodieken om volledige reproduceerbaarheid te garanderen.

B2: Architectuur en AI-technieken selecteren

S/T: Selecteren van een architectuur voor tekstgeneratie (Fine-tuning vs RAG). **A:** Ik heb de feedback uit Blok 1 (BEW-007) toegepast door te starten met een **nulmodel/baseline** (DummyRegressor in M&T) om de toegevoegde waarde van complexere modellen aan te tonen (BEW-201). **R:** De keuze voor de architectuur

was in de M&T opdrachten goed onderbouwd (BEW-202). **T:** Het juist kiezen van het nulmodel en onderscheid met een baseline model is iets wat ik in de toekomst kan meenemen.

B3: Systematische modelontwikkeling

S/T: Iteratieve ontwikkeling van modellen met aandacht voor toeval en hyperparameters. **A:** In opdracht 1b heb ik systematisch 5 runs per experiment uitgevoerd om rekening te houden met toeval (BEW-201). Bij opdracht 2a heb ik iteraties uitgevoerd op dropout-rates en learning rates, onderbouwd met literatuur (BEW-202). Ook in het groepsproject heb ik geholpen met de modeliteraties voor de verschillende modellen die zijn geevalueerd. **R:** Hoewel de aanpak in de individuele opdrachten systematisch was en ook goed beoordeeld, ontbrak in het groepsproject de documentatie van de uitvoering op Git, waardoor het niet beoordeeld kon worden (BEW-205). **T:** Voor de finale inlevering zorg ik dat alle experiment-logs en model-iteraties volledig in de repository staan.

C. EVALUEREN EN MONITOREN

C1: Maatschappelijke impact evalueren

S/T: De consequenties van geautomatiseerde samenvattingen voor de burger evalueren. **A:** Ik heb bijgedragen aan de ethische verantwoording in het rapport. We hebben gekeken naar welke requirements wel/niet behaald zijn en wat dit betekent voor burgers met een taalachterstand. **R:** De feedback was positief over de slotdiscussie, maar adviseerde om de impact nog concreter te maken door kosten en baten voor de organisatie en maatschappij af te wegen (BEW-204). **T:** In de eindversie zal dit door een groepsgenoot worden overgenomen, dus op het moment kan ik dit niet koppelen aan een stap in de toekomst, maar ik zal proberen de slotdiscussie te controleren en verbeteren indien nodig.

C2: Kwaliteitscriteria toepassen

S/T: Het model evalueren op meer dan alleen standaard statistieken. **A:** In lijn met de transfer uit Blok 1 heb ik naast MAE (voor regressie) nu ook gekeken naar complexe maatstaven zaols de interpreteerbaarheid van coëfficiënten en metrieken zoals BLEU voor het project (BEW-201, BEW-205). Ook hebben we 'duurzaamheid' (resource demand) meegenomen als criterium voor het project (BEW-205). **R:** De evaluatie in M&T was 'Boven niveau'. In de technische review van het project werd opgemerkt dat de verantwoording voor de keuze van BLEU sterker kon en dat we voorzichtig moeten zijn met synthetische data als ground-truth (BEW-205). **T:** Ik ga de selectie van evaluatiemetrieken beter funderen met literatuur die de correlatie tussen deze metrieken en menselijke beoordeling beschrijft. Als update heb ik alvast dat dit is toegepast en de metrieken zijn beter gefundeerd en synthetische data is niet meer als ground-truth gebruikt.

C3: Prototype testing met stakeholders

S/T: Het prototype valideren in de context van de gebruiker. **A:** We hebben tests uitgevoerd met een paper prototype voor de medewerkers van de OVER-gemeenten. **R:** De tests met de medewerkers van de gemeente waren positief maar de meeste medewerkers konden niet makkelijk de juiste sectie vinden om een samenvatting te kunnen maken. **T:** De feedback nemen we op in ons volgende prototype, voor mij houdt dat in dat de flows beter worden gedefinieerd op basis van de feedback van de medewerkers.

D. ZELFSTURING

D1: Feedback verzamelen en verwerken

S/T: Het systematisch integreren van feedback uit Blok 1 in de werkzaamheden van Blok 2. **A:** Ik heb een actieve houding aangenomen door de 'Transfer'-punten uit Blok 1 als checklist te gebruiken voor Blok 2. Voorbeelden zijn:

- Het voorkomen van data leakage (B1).
- Het introduceren van baselinemodellen (B2).
- Het toevoegen van bronvermeldingen aan requirements (A2). **R:** Dit logboek toont aan dat ik de feedback niet alleen heb ontvangen, maar ook heb omgezet in verbeterde resultaten, zoals te zien is in de 'Boven niveau' scores voor B1 en C2 in de recente M&T feedback (BEW-202) en de technische review (BEW-205). **T:** Voor het afstudeerproject is het belangrijk om een feedback-loop te implementeren. Dit kan tijdens de wekelijke meeting gebeuren.

D2: Onderzoek opzetten en rapporteren

S/T: Een PvA schrijven voor een artikel over PDF parsing en dit presenteren. **A:** Ik heb een PvA opgesteld volgens de IEEE-standaarden. In de presentatie op het symposium heb ik de methodiek mondeling toegelicht. **R:** De presentatie was 'Zeer goed georganiseerd' met sterke vaardigheden (BEW-207). Het PvA werd echter kritisch beoordeeld op ambitie en de 'gap' in literatuur (BEW-206). Het werd als te beschrijvend ervaren. **T:** Voor het uiteindelijke artikel ga ik de focus verleggen van een algemene tutorial naar een systematische vergelijking van parsing-tools voor specifieke downstream NLP-taken.

D3: Robuuste onderzoeksmethoden kiezen

S/T: Kiezen van valide methoden voor het PDF parsing onderzoek. **A:** Ik heb geprobeerd een systematische aanpak te kiezen voor het verwerken van gemeentelijke documenten. **R:** In het schriftelijke PvA was de methodiek nog 'te beperkt' en miste een systematische aanpak voor het tonen van resultaten (BEW-206). Op de presentatie werd D3 als 'Op niveau' beoordeeld omdat ik vragen over de methodiek goed kon beantwoorden sinds ik de feedback van het PvA had toegepast (BEW-207). **T:** Ik ga onderzoek doen om een set aan tools te evalueren in een evaluatiematrix en zo op basis van de gegeven criteria lezers de juiste tool kunnen laten selecteren voor hun downstream NLP-taak.