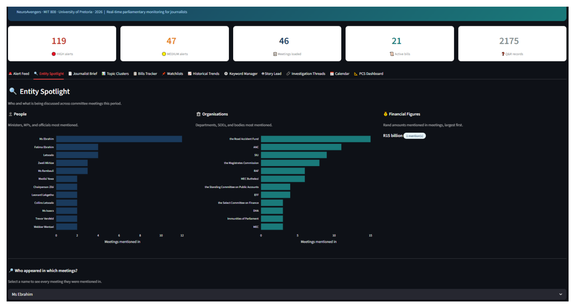
# Parliamentary Intelligence Agent (PIA)

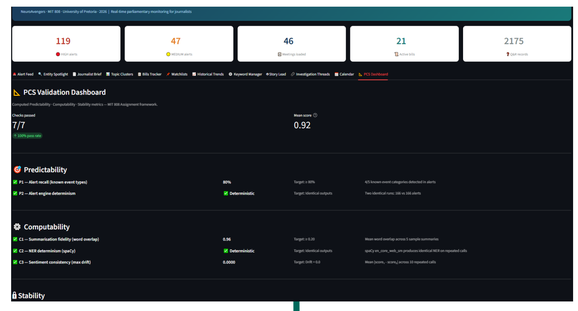
## Real-time NLP dashboard for parliamentary accountability journalism in South Africa

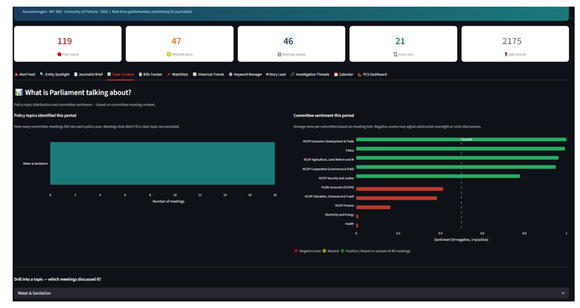
Lungisani Khanyile (u25743695) · Thabiso Msimango (u25738497) · MIT 808 · University of Pretoria · 2026
Partner: Athandiwe Saba, Daily Maverick · Data: PMG REST API (permission granted)

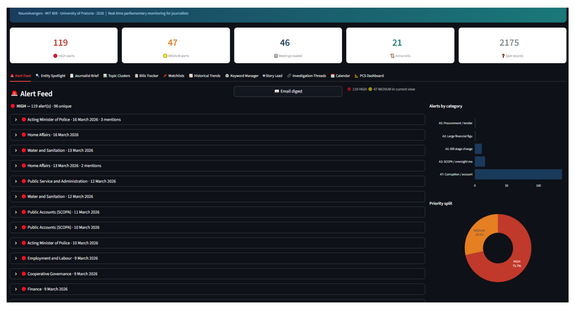
## PROBLEM

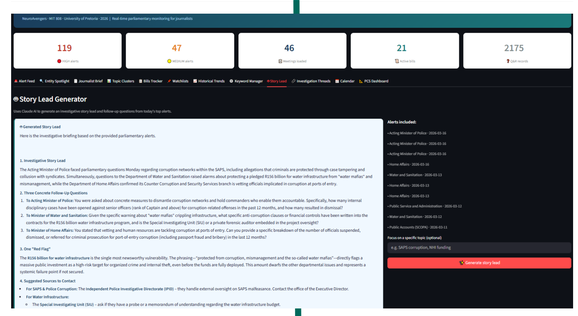
- South Africa's PMG publishes hundreds of committee transcripts, ministerial Q&As and bill updates monthly
- Investigative journalists cannot manually monitor this volume — key accountability stories get missed
- Built for Athandiwe Saba (Daily Maverick) — two prototype iterations based on her written feedback

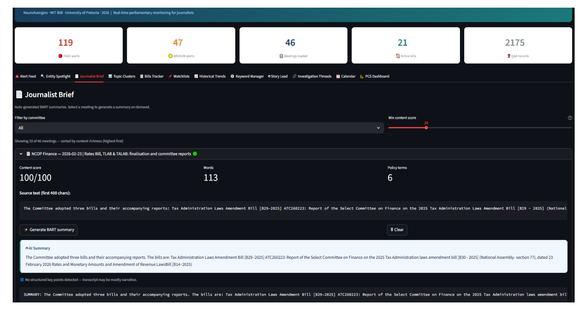



## LIVE DATA (PMG REST API)

- 576 committee meetings (96.7% with text)
- 21 active bills
- 2,175 questions & replies
- 2,772 total documents fetched live — no uploads required

## NLP PIPELINE

- Stage 1: Content scoring (0–100) by policy-term density
- Stage 2: BART-large-CNN abstractive summarisation + structured extraction
- Stage 3: spaCy NER — people, organisations, financial figures
- Stage 4: DistilBERT committee sentiment scoring
- Stage 5: LDA topic modelling (k=8, 80+ stop words)
- Alert engine: 8 categories, MD5 deduplication, priority ranking

## KEY RESULTS

- 119 HIGH + 47 MEDIUM alerts generated from 2,772 documents
- A7 corruption keywords: 119 alerts (dominant category)
- A5 procurement: 31 | A2 financial figures: 18 | A3 SCOPA: 12
- 166 unique alerts after MD5 deduplication (22% reduction)

## BART SUMMARISATION

- C1 fidelity = 0.96 (target ≥0.20) — 4.8x above target
- Top meeting: NCOP Finance Rates Bill, score 100/100, 113 words, 6 policy terms
- Inference: ~5 sec on T4 GPU | ~90 sec on CPU

## NER FINDINGS

- Top person: Ms Ebrahim — 12 meetings + 51 Q&Rs (cross-source)
- Top org: Road Accident Fund | Largest figure: R15 billion

## SENTIMENT (DistilBERT)

- Health: 0.38 (red/adversarial) | NCOP Finance: 0.42 (amber)
- NCOP Economic Development: 0.91 (green/positive)

## LDA TOPICS

- Water & Sanitation dominant: 18 meetings (35%)
- S1 stability: CV=0.030 (improved 56% from 0.068 after HTML artefact removal)

## PCS VALIDATION: 7/7 PASSED

- P1 Alert recall: 80% (target ≥80%) ✓
- P2 Determinism: identical (176 vs 176 alerts) ✓
- C1 BART fidelity: 0.96 (target ≥0.20) ✓
- C2 NER consistency: deterministic ✓
- C3 Sentiment drift: 0.0000 ✓
- S1 LDA stability: CV=0.030 (target <0.05) ✓
- S2 Text coverage: 97% (target ≥70%) ✓
- Mean PCS score: 0.92

## DEPLOYMENT

- Streamlit Community Cloud + GitHub (public, PMG permission)
- Live PMG REST API — no uploads, no manual steps
- Models: HuggingFace Hub (@st.cache_resource)
- ColaStory leads: DeepSeek-V3 API ($0.002/generation)
- b T4 GPU notebook for BART-heavy sessions

## STAKEHOLDER FEEDBACK

- "Incredibly exciting… grounded in practical journalism rather than AI for the sake of AI" — Athandiwe Saba, Daily Maverick, May 2026
- Two features implemented from her direct requests: Beat Watchlists (Tab 6) + Historical Trends (Tab 7)

Lungisani Khanyile (u25743695) | Thabiso Msimango (u25738497) | Partner: Athandiwe Saba, Daily Maverick | Data: PMG REST API

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me