SocialPulse Monastir — Sentiment, sujets & détection d'événements

Build a near real-time social media monitoring system that:

- Tracks public sentiment (positive/negative/neutral) and emerging topics from noisy, multilingual social media posts (mainly Twitter/X, Facebook public pages…)
- Detects local events (e.g., traffic jams, tourism…) by spotting anomalies or bursts in topic/sentiment signals.
- Generates prioritized alerts for local stakeholders.

A. Multilingual & Noisy Text Preprocessing

-Transliteration normalization

Camel Tools : Has a transliterator for Arabic ↔ Latin

Qalsadi : Arabic morphological analyzer (works best with Arabic script)

-Emoji handling

Replace with sentiment tokens

-Language detection

Use **fastText** or **CLD3** (Google's Compact Language Detector)

B. Sentiment Analysis (Multilingual)

- Bronze: Lexicon-based (e.g., **VADER** for English/French, **ArSentD-LEV** for Arabic dialects).
- Silver/Gold: Fine-tune multilingual **BERT** (e.g., mBERT, XLM-R) on Tunisian/French sentiment data.
    - Datasets:
        - ArSAS (Arabic Sentiment Analysis)
        - SemEval-2017 Task 4 (Arabic sentiment)
        - Collect local Monastir-relevant tweets (with hashtags like #Monastir, #المنستير)

C. Topic Modeling & Semantic Clustering

- Bronze: **LDA** or **NMF** on **TF-IDF** vectors (after stopword removal).

- Silver: Use sentence embeddings (e.g., LaBSE, paraphrase-multilingual-MiniLM) → cluster with HDBSCAN (handles noise well).
- Gold: Dynamic topic models (e.g., BERTopic) + topic coherence (NPMI).

D. Event Detection (Temporal Anomaly Detection)

- Change-point detection:
    - Bayesian Online Change Point Detection (Adams & MacKay, 2007)
    - Twitter's AnomalyDetection R package (ported to Python)
- Spike detection: Monitor topic/sentiment time series → flag when Z-score > 3 or use EWMA (Exponentially Weighted Moving Average).
- Gold: Build a temporal knowledge graph linking entities (e.g., "Monastir airport") + topics + sentiment.

E. Alert Prioritization

- Confidence score (from model)
- Volume spike (number of posts)
- Sentiment polarity shift (e.g., sudden negativity about "electricity")
- Entity relevance (using local gazetteer: hospitals, roads, beaches)

Week 1: Setup & Data

- Define data sources (Twitter API, simulated dump)
- Build local gazetteer (landmarks, roads, institutions in Monastir)
- Annotate small dataset (sentiment + events)

Python, Twitter API v2, GeoNames, manual annotation (Label Studio)

Week 2-3: Bronze Pipeline

- Streaming ingestion (Your system **listens continuously** to a data source.)
(Kafka/Pulsar or simple file watcher)
- Preprocessing: emoji (demoji library)→ text, Darija translit normalization
- Language detection(langdetect, fastText)
- Lexicon-based sentiment (VADER + Arabic lexicons)
- LDA/NMF topic modeling
- Simple Streamlit dashboard (time series + word clouds)

emoji,textblob,gensim,scikit-learn,streamlit

Week 3–4: Silver Upgrade (Advanced Sentiment Analysis + Semantic Topics & Clustering)

- Fine-tune**XLM-R**on local sentiment data
- Generate sentence embeddings → semantic clustering (HDBSCAN)
- Improve topic coherence
- Add alert logic

HuggingFace Transformers,sentence-transformers,hdbscan

Week 5–6: Gold Features

- Bayesian change-point detection on topic/sentiment streams
- Entity linking (match “مستشفى” → “Monastir Regional Hospital”)
- Build topic graph (nodes = topics, edges = co-occurrence)

ruptures(Python lib), spaCy + custom NER, NetworkX

Week 7–8: Platinum & Evaluation

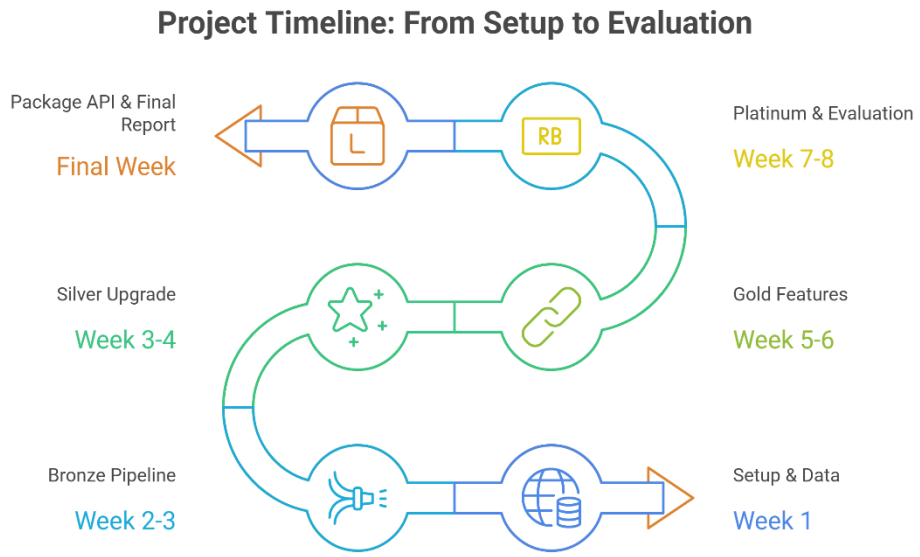
- RAG-style QA: “What happened in Monastir last Tuesday?” → retrieve relevant posts
- Counterfactual robustness tests (e.g., perturb emojis, swap dialects)
- Freeze test sets, compute F1, NPMI, Precision@k, lead time

LangChain, FAISS, custom evaluation scripts

Final Week

- Package API (FastAPI)
- Final Streamlit dashboard
- report

FastAPI, MLflow, Streamlit, GDPR checklist



Recommended Reading

Essential Papers

1. **"Sentiment Analysis for Arabic on Social Media" (2021)** - Comprehensive survey
2. **"BERTopic: Neural Topic Modeling"** - Modern topic modeling approach
3. **"Event Detection in Twitter" (WWW 2012)** - Classic event detection methods
4. **"Bayesian Online Changepoint Detection" (2007)** - Mathematical foundations

Practical Guides

- Hugging Face Course (multilingual NLP module)
- "Speech and Language Processing" (Jurafsky & Martin) - Chapters 4, 6, 21
- Fast.ai NLP course

Tunisian/Maghrebi NLP Resources

- **TUNIZI** dataset (Tunisian Arabic sentiment)
- **Maghrebi Arabic Dialect corpus**
- Research from ANLP workshops (Arabic NLP)