

SentiNela: assistente clínico conversacional e vigilância sindrômica com modelo de linguagem e previsão de tendências

Leonardo Arazo de Oliveira Araújo¹

¹ Universidade Federal de São Paulo (UNIFESP). (arazo.leonardo@unifesp.br).

Resumo: A vigilância sindrômica utiliza dados de atendimentos para detectar surtos e tendências epidemiológicas em tempo próximo ao real. Este trabalho apresenta o SentiNela, sistema que integra assistente clínico conversacional baseado em modelo de linguagem, pipeline de vigilância sindrômica, previsão de tendências e detecção de anomalias, objetivando triagem e categorização de queixas em português sem diagnóstico médico, apenas categorização sindrômica e orientação sobre onde buscar atendimento. Na camada conversacional, MedGemma (Vertex AI), ajustado por prompt, acolhe o usuário, coleta sintomas, duração e faixa etária e indica onde tratar-se; a extração estruturada ocorre em background, exibindo-se só a conversa. O pipeline reúne inferência no Vertex AI, extração de características clínicas, armazenamento e agregação temporal. Na previsão e vigilância: EMA e suavização de Holt (até 7 dias por categoria), regressão linear para tendência, Z-score e taxa de crescimento para anomalias; um score composto classifica atenção por categoria e alimenta alertas a gestores com notificação opcional por webhook ou Telegram. A metodologia compreende oito categorias sindrômicas e níveis de severidade, inferência com histórico de conversa, agregação temporal, módulo TrendPredictor, detector de anomalias e API REST para chat, previsões e analytics. O piloto (janeiro/2026, três semanas, duas clínicas, aproximadamente 70 conversas com consentimento) apresentou completude de extração de aproximadamente 97% para categoria, 94% severidade, 91% duração e 93% sintomas. Séries diárias esparsas (~3 interações/dia) restringiram previsão robusta a 3–5 categorias; tendência majoritariamente estável (~75%); score médio por categoria de 9/100 e agregado de 18/100; 0–1 evento anômalo leve, sem surto epidemiológico. Para validação retrospectiva do pipeline sobre série epidemiológica pública, o mesmo método (EMA $\alpha=0,30$; Holt-Winters $\alpha=0,30$, $\beta=0,10$, $\gamma=0,20$, sazonalidade semanal) foi aplicado à série SIVEP-Gripe 2024 de São Paulo (18.017 registros, 366 dias, média de 49,2 casos/dia). Identificaram-se 15 anomalias por Z-score acima de $2,0\sigma$ e 16 por taxa de crescimento acima de 50%, coerentes com os picos sazonais de SRAG documentados. Em backtest robusto (média de 5 holdouts de 7 dias, fevereiro a outubro, excluindo dezembro por atraso de notificação), o Holt-Winters obteve MAE médio de 18,1 casos/dia e RMSE de 21,0 casos/dia, reduzindo em 46% o erro frente ao Holt simples (MAE 33,7); o MAPE de 48% reflete o perfil heterocedástico da série, sendo o MAE absoluto o indicador principal. As principais limitações são volume reduzido do piloto, séries curtas, dependência de prompts e generalização restrita a duas unidades. Os próximos passos incluem ampliar duração e número de unidades, calibrar prospectivamente limiares e pesos do score, realizar auditoria da extração e endereçar requisitos de produção (LGPD, segurança) e integrações institucionais. O SentiNela é um protótipo de apoio à vigilância sindrômica e orientação ao cidadão, passível de validação e expansão em saúde pública.

Palavras-chave: Saúde, Inteligência Artificial, Séries Temporais.

SentiNela: Conversational Clinical Assistant and Syndromic Surveillance with Language Model and Trend Forecasting

Leonardo Arazo de Oliveira Araújo¹

¹ Federal University of São Paulo (UNIFESP). (arazo.leonardo@unifesp.br).

Abstract: Syndromic surveillance uses healthcare encounter data to detect outbreaks and epidemiological trends in near real time. This work presents SentiNela, a system that integrates a conversational clinical assistant based on a large language model, a syndromic surveillance pipeline, trend forecasting, and anomaly detection. Its objective is the triage and categorization of complaints in Portuguese without medical diagnosis, providing only syndromic categorization and guidance on where to seek care. In the conversational layer, MedGemma (Vertex AI), adjusted via prompt engineering, welcomes the user, collects symptoms, duration, and age group, and indicates appropriate care settings; structured extraction occurs in the background, exposing only the conversation to the user. The pipeline integrates Vertex AI inference, clinical feature extraction, storage, and temporal aggregation. For forecasting and surveillance, the system employs exponential moving average (EMA) and Holt smoothing (up to 7 days per category), linear regression for trend estimation, Z-score and growth rate for anomaly detection; a composite score classifies alert levels by category and feeds manager notifications via optional webhook or Telegram integration. The methodology encompasses eight syndromic categories and severity levels, inference with conversation history, temporal aggregation, a TrendPredictor module, an anomaly detector, and a REST API for chat, forecasting, and analytics. A pilot study (January 2026, three weeks, two clinics, approximately 70 conversations with informed consent) yielded extraction completeness of approximately 97% for category, 94% for severity, 91% for duration, and 93% for symptoms. Sparse daily series (~3 interactions/day) restricted robust forecasting to 3–5 categories; trends were predominantly stable (~75%); mean score per category was 9/100 and aggregate 18/100; 0–1 mild anomalous events were observed, with no epidemiological outbreak detected. For retrospective validation of the pipeline on a public epidemiological series, the same method (EMA $\alpha=0.30$; Holt-Winters $\alpha=0.30$, $\beta=0.10$, $\gamma=0.20$, weekly seasonality) was applied to the SIVEP-Gripe 2024 series for the city of São Paulo (18,017 records, 366 days, mean 49.2 cases/day). Fifteen anomalies were identified by Z-score above 2.0σ and 16 by growth rate above 50%, consistent with documented seasonal SARI peaks. In robust backtesting (mean of 5 seven-day holdouts, February through October, excluding December due to notification lag), Holt-Winters achieved a mean MAE of 18.1 cases/day and RMSE of 21.0 cases/day, reducing error by 46% relative to simple Holt smoothing (MAE 33.7); the MAPE of 48% reflects the heteroscedastic profile of the series, with absolute MAE as the primary indicator. Key limitations include the reduced volume of the pilot, short time series, prompt dependency, and generalizability restricted to two units. Next steps include extending duration and number of participating units, prospectively calibrating score thresholds and weights, conducting extraction audits, and addressing production requirements (LGPD compliance, security) and institutional

integrations. SentiNela is a prototype supporting syndromic surveillance and citizen guidance, amenable to validation and expansion in public health.

Keywords: Health, Artificial Intelligence, Time Series.