

UNIVERSIDADE VIRTUAL DO ESTADO DE SÃO PAULO

BACHARELADO EM CIÊNCIA DE DADOS

CINTIA IZUMI SHINODA

RELATÓRIO FINAL DE ESTÁGIO CURRICULAR OBRIGATÓRIO

São Paulo - SP

2026

UNIVERSIDADE VIRTUAL DO ESTADO DE SÃO PAULO

BACHARELADO EM CIÊNCIA DE DADOS

CINTIA IZUMI SHINODA

RELATÓRIO FINAL DE ESTÁGIO CURRICULAR OBRIGATÓRIO

Relatório desenvolvido como requisito para
aprovação no componente curricular Estágio
Curricular Obrigatório no curso Bacharelado
em Ciência de Dados na Universidade
Virtual do Estado de São Paulo (UNIVESP).

São Paulo - SP

2026

RELATÓRIO FINAL DE ESTÁGIO CURRICULAR OBRIGATÓRIO

APRESENTAÇÃO

Nome: Cintia Izumi Shinoda

RA: 

Curso: Bacharelado em Ciência de Dados

Nome do Supervisor Responsável pela Empresa: Ademir Stefanski Junior

Empresa/Entidade: MUSAKANON SERVIÇOS - SOCIEDADE UNIPessoal LTDA

Ramo de Atividade: Desenvolvimento e licenciamento de programas de computador customizáveis

Período de Estágio: de 02/02/2026 a 27/02/2026 (Exceto 16/02, 17/02 e 18/02/2026)

Quantidade total de horas cumpridas: 100 horas

São Paulo, 27 de fevereiro de 2026.



Ademir Stefanski Junior (Supervisor do Estágio)



Cintia Izumi Shinoda (Estagiária)

LISTA DE ILUSTRAÇÕES

Figura 1 - Dashboard: Métricas Gerais e Distribuição de Sentimento.....	18
Figura 2 - Dashboard: <i>Heatmap</i> : Tópico x Sentimento	19
Figura 3 - Dashboard: Top tópicos por volume.....	20
Figura 4 - Dashboard: <i>Heatmap</i> temporal: Tópicos por Sessão.....	20
Figura 5 - Dashboard: Distribuição de confiança por sentimento.....	21
Figura 6 - Dashboard: Explorador de segmentos	21
Figura 7 - Dashboard: Nuvem de palavras.....	22

LISTA DE TABELAS

Tabela 1 - Tecnologias empregadas em cada fase do <i>pipeline</i>	11
Tabela 2 - <i>Benchmark</i> comparativo para seleção do Modelo de Sentimento.....	12

SUMÁRIO

1 INTRODUÇÃO.....	7
1.1 Produtos/Serviços Desenvolvidos pela Empresa.....	7
1.2 Importância do Estágio para a Formação Profissional	7
2 ANÁLISE DA ESTRUTURA ORGANIZACIONAL	9
2.1 Objetivos da Empresa e Função Social.....	9
2.2 Setor Econômico	9
3 CARACTERIZAÇÃO DA ÁREA DE ATUAÇÃO PROFISSIONAL	11
3.1 Principais Funções da Área.....	11
3.2 Metodologias Utilizadas no Desenvolvimento das Atividades.....	11
3.2.1 Stack Tecnológico	11
3.2.2 Seleção do Modelo de Sentimento.....	12
3.2.3 Extração de Tópicos.....	13
3.3 Atividades Desenvolvidas.....	13
3.3.1 1º ao 5º Dia: Configuração e Coleta de Dados	13
3.3.2 6º ao 12º Dia: Análise e Visualização	14
3.3.3 13º ao 17º Dia: Documentação e Entrega	14
3.4 Relação das Disciplinas Cursadas com as Atividades do Estágio	15
3.4.1 Estatística e Probabilidade	15
3.4.2 Banco de Dados	15
3.4.3 Aprendizado de Máquinas	15
3.4.4 Processamento de Linguagem Natural.....	15
3.4.5 Visualização Computacional.....	15
3.4.6 Projetos e Métodos para a Produção do Conhecimento.....	15
3.4.7 Engenharia de Software	15
4 CONCLUSÃO E RECOMENDAÇÕES.....	16
4.1 Conclusão.....	16
4.2 Sugestões	16
4.2.1 Para o Projeto.....	16
4.2.2 Para a Empresa.....	17
4.3 Resultado	18
5 REFERÊNCIAS.....	23
ANEXO - FICHA DE PRESENÇA.....	25

1 INTRODUÇÃO

A Musakanon Serviços - Sociedade Unipessoal LTDA, registrada sob o CNPJ

é uma microempresa constituída em 29 de agosto de 2025, sediada em São Paulo, SP. Seu objetivo abrange a reparação e manutenção de computadores e equipamentos periféricos, além do desenvolvimento e licenciamento de programas de computador customizáveis.

O presente estágio foi realizado na área de Ciência de Dados aplicada à análise de discurso legislativo. O projeto consistiu na construção de um *pipeline* de Processamento de Linguagem Natural (NLP), capaz de coletar, transcrever, classificar sentimentos e extrair tópicos de sessões plenárias da Câmara Municipal de São Paulo, disponibilizando os resultados em um *dashboard* interativo.

O volume de sessões plenárias produzidas pela Câmara Municipal excede a capacidade de acompanhamento tanto da sociedade civil quanto da imprensa. O problema motivador deste projeto foi transformar centenas de horas de discurso parlamentar em informação estruturada e acessível. A solução encontrada foi a construção de um *pipeline* automatizado que parte do áudio bruto dos vídeos disponíveis no YouTube e entrega *insights* analíticos em tempo real.

1.1 PRODUTOS/SERVIÇOS DESENVOLVIDOS PELA EMPRESA

A Musakanon Serviços atua no desenvolvimento de soluções de software customizáveis, com foco em automação de processos e análise de dados. Dentre os serviços oferecidos, destacam-se: desenvolvimento web, consultoria em ciência de dados e manutenção de software.

1.2 IMPORTÂNCIA DO ESTÁGIO PARA A FORMAÇÃO PROFISSIONAL

O estágio em Ciência de Dados proporcionou a oportunidade de aplicar, em contexto real, os conhecimentos adquiridos ao longo do curso da UNIVESP. O projeto exigiu competências em coleta e ingestão de dados (*web scrapping*), processamento de áudio (*Speech-to-Text*), pré-processamento textual (NLP), modelagem de *Machine Learning* (análise de sentimento e modelagem de tópicos) e visualização de dados (*dashboard* interativo).

Além das competências técnicas, o estágio permitiu desenvolver habilidades de gestão de projeto, tomada de decisão baseada em dados, comunicação de resultados técnicos e pensamento sistêmico.

2 ANÁLISE DA ESTRUTURA ORGANIZACIONAL

A Musakanon Serviços é uma sociedade empresária limitada unipessoal, classificada como microempresa, constituída em 29 de agosto de 2025. Sua estrutura é enxuta, com gestão centralizada no sócio administrador Ademir Stefanski Junior, e com atuação focada em desenvolvimento de software: desenvolvimento *web* e soluções de dados.

Por ser uma microempresa de estrutura organizacional horizontal, não há departamentos formais ou hierarquias intermediárias. As decisões técnicas e estratégicas são tomadas diretamente pelo sócio administrador em colaboração com profissionais envolvidos em cada projeto. Essa característica foi determinante durante o estágio, pois permitiu participação ativa em todas as fases do projeto, desde a definição do escopo e escolha das tecnologias até a entrega do projeto final, sem barreiras burocráticas comuns em organizações de maior porte.

2.1 OBJETIVOS DA EMPRESA E FUNÇÃO SOCIAL

A empresa tem como objetivo oferecer soluções tecnológicas acessíveis, com ênfase em desenvolvimento *web* e análise de dados. No contexto do projeto desenvolvido durante o estágio, a função social é especialmente relevante: a ferramenta contribui diretamente para a transparência, permitindo que cidadãos, jornalistas e organizações da sociedade civil acompanhem de forma estruturada os debates legislativos da maior câmara municipal do país. Ao disponibilizar análises automatizadas de sentimento e tópicos que mais estão presentes nas sessões plenários, o projeto materializa o compromisso da empresa com o uso da tecnologia como instrumento de cidadania e responsabilidade governamental.

2.2 SETOR ECONÔMICO

A empresa atua no setor de Tecnologia da Informação, especificamente nos segmentos de desenvolvimento de *software* e análise de dados. O mercado brasileiro de TI movimenta cifras expressivas e apresenta crescimento contínuo, impulsionado pela transformação digital em organizações públicas e privadas.

A demanda por soluções baseadas em dados se intensificou nos últimos anos, criando oportunidades para microempresas que oferecem serviços especializados e customizáveis.

Dentro desse cenário, o nicho de NLP aplicado ao setor público ainda é pouco explorado no Brasil. Embora existam iniciativas de dados abertos em câmaras municipais e assembleias legislativas, a maior parte do conteúdo legislativo permanece em formato não-estruturado, o que representa uma barreira significativa para o acompanhamento do cidadão. Essa lacuna posiciona empresas com competência em Ciência de Dados e Processamento de Linguagem Natural como agentes potenciais de transformação na transparência democrática.

3 CARACTERIZAÇÃO DA ÁREA DE ATUAÇÃO PROFISSIONAL

O estágio foi realizado na área de Ciência de Dados, com foco em Processamento de Linguagem Natural (NLP) aplicado à análise de discurso político.

Esta é uma subárea da Inteligência Artificial que combina Linguística Computacional, Estatística e Aprendizado de Máquina para extrair informações estruturadas de texto e áudio não estruturados.

3.1 PRINCIPAIS FUNÇÕES DA ÁREA

As atividades de Ciência de Dados no contexto deste estágio envolveram todo o ciclo analítico: definição do problema, coleta de dados, processamento de limpeza, modelagem, avaliação de resultados e comunicação por meio de visualizações.

Especificamente, as funções desempenhadas incluíram: ingestão automatizada de dados multimídia, transcrição de áudio em larga escala, análise de sentimento com modelos *Transformer*, extração não-supervisionada de tópicos e construção de *dashboard* interativo.

3.2 METODOLOGIAS UTILIZADAS NO DESENVOLVIMENTO DAS ATIVIDADES

O desenvolvimento do projeto seguiu a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*), adaptada para o contexto de NLP. As seis fases foram executadas ao longo das 100 horas de estágio.

3.2.1 STACK TECNOLÓGICO

A tabela a seguir detalha as tecnologias empregadas em cada fase do *pipeline*:

Tabela 1 - Tecnologias empregadas em cada fase do *pipeline*

Fase	Tecnologia	Função
Coleta	yt-dlp	Download do áudio das sessões plenárias do YouTube

Transcrição	faster-wisper (large-v3)	Speech-to-Text com modelo Whisper da OpenAI, otimizado com CTranslate2
Pré-processamento	spaCy (pt_core_news_lg)	Tokenização, lematização, remoção de <i>stopwords</i> , filtragem de segmentos
Sentimento	cardiffnlp/xlm-roberta	Classificação de sentimento (positivo, neutro ou negativo) com <i>Transformer</i> multilíngue
Tópicos	BERTopic	Extração não-supervisionada de tópicos com <i>embeddings</i> + UMAP + HDBSCAN
Armazenamento	DuckDB + Parquet	Catálogo analítico com consultas SQL sobre arquivos colunares
<i>Dashboard</i>	Streamlit + Plotly	Interface <i>web</i> interativa com 9 visualizações e filtros em tempo real
Versionamento	Git	Controle de versão do código e documentação

Fonte: Elaborado pela autora

3.2.2 SELEÇÃO DO MODELO DE SENTIMENTO

A escolha do modelo de análise de sentimento foi precedida por um *benchmark* comparativo com três modelos pré-treinados, avaliados sobre seis frases representativas do discurso parlamentar (incluindo 2 frases neutras procedimentais):

Tabela 2 - *Benchmark* comparativo para seleção do Modelo de Sentimento

Modelo	Acertos	Neutros	Confiança
nlptown/bert-multilingual	4/6	0/2	33-93%

lxyuan/distilbert-multilingual	4/6	0/2	86-96%
cardiffnlp/xlm-roberta	5/6	2/2	50-96%

Fonte: Elaborado pela autora

O modelo cardiffnlp/twitter-xlm-roberta-base-sentiment foi selecionado por ser o único a acertar as duas frases neutras. Em sessões legislativas onde aproximadamente 73% do discurso é procedimental (neutro), a capacidade de distinguir conteúdo opinativo de conteúdo processual é determinante para a qualidade da análise.

3.2.3 EXTRAÇÃO DE TÓPICOS

A extração de tópicos utilizou BERTopic com *embeddings* do modelo paraphrase-multilingual-MiniLM-L12-v2, redução de dimensionalidade com UMAP (random_state=42, n_components=5, n_neighbors=15, min_dist=0.0, metric cosine) e clusterização com HDBSCAN (min_cluster_size=50, min_samples=5). O parâmetro min_cluster_size foi calibrado experimentalmente: com 10, o modelo produziu 174 tópicos fragmentados; com 50, produziu 43 tópicos coerentes.

3.3 ATIVIDADES DESENVOLVIDAS

O estágio foi organizado em 17 dias úteis, totalizando 100 horas de atividade distribuídas em quatro semanas (excluindo o período de Carnaval de 16 a 18 de fevereiro). As atividades descritas a seguir em ordem cronológica.

3.3.1 1º AO 5º DIA: CONFIGURAÇÃO E COLETA DE DADOS

Configuração do ambiente de desenvolvimento (Python 3.11, venv, Git). Download de 6 sessões plenárias via yt-dlp, totalizando 12 horas e 28 minutos de áudio (8 GB em WAV). Catalogação das sessões em DuckDB com metadados (duração, data e número de segmentos). Transcrição completa com faster-whisper (modelo large-v3, device=cpu, compute_type=int8, beam_size=5, VAD filter ativado). Tempo de processamento: 7 horas e 20 minutos para 12

horas e 28 minutos de áudio (1,9x tempo real). Resultado: 11.435 segmentos transcritos, 68.133 palavras.

3.3.2 6º AO 12º DIA: ANÁLISE E VISUALIZAÇÃO

Pré-processamento com spaCy (pt_core_news_lg): filtragem de vinhetas de abertura, remoção de segmentos curtos (menos do que 3 palavras), lematização preservando entidades nomeadas (como nomes). Resultado: 9.239 segmentos limpos (81% de aproveitamento).

Análise de sentimento com cardiffnlp/twitter-xlm-roberta-base-sentiment sobre os 9.239 segmentos. Distribuição global: 10% positivo, 73% neutro, 17% negativo, com confiança média de 71%.

Extração de 43 tópicos com BERTopic. O tópico genérico (#0, discurso procedimental) concentrou 30% dos segmentos após refinamento com *seed* fixa.

Análise de sub-clusterização do tópico #0 revelou 31 subtemas ocultos (orçamento, crime, políticas públicas).

Construção do *dashboard* Streamlit com 9 seções: métricas gerais, distribuição de sentimento, sentimento por sessão, *heatmap* Tópico x Sentimento (top 25), tópicos mais frequentes, *heatmap* temporal, distribuição de confiança, explorador de segmentos e nuvem de palavras.

Validação qualitativa com amostra de 15 segmentos (5 por sentimento, random_state=42), Acurácia estimada: 80-93%. Erros concentrados em frases procedimentais ambíguas.

3.3.3 13º AO 17º DIA: DOCUMENTAÇÃO E ENTREGA

Redação do relatório final do estágio e criação de apresentação para os *stakeholders* do projeto.

Desenvolvimento de *script* de teste *end-to-end* (20 verificações automatizadas).

Refinamento do README e da documentação do projeto.

Teste completo do *pipeline* e preparação para entrega

3.4 RELAÇÃO DAS DISCIPLINAS CURSADAS COM AS ATIVIDADES DO ESTÁGIO

Diversas disciplinas do curso de Ciência de Dados da UNIVESP foram mobilizadas durante o estágio. A seguir, apresenta-se as disciplinas e as atividades desenvolvidas relacionadas:

3.4.1 ESTATÍSTICA E PROBABILIDADE

Avaliação de distribuições de confiança do modelo, métricas de acurácia, amostragem para validação qualitativa.

3.4.2 BANCO DE DADOS

Modelagem e armazenamento em Parquet e DuckDB, consultas analíticas com SQL

3.4.3 APRENDIZADO DE MÁQUINAS

Seleção e avaliação de modelos de sentimento (XML-RoBERTa) extração de tópicos (BERTopic)

3.4.4 PROCESSAMENTO DE LINGUAGEM NATURAL

Tokenização, lematização, remoção de *stopwords* com spaCy, *embeddings* multilíngues

3.4.5 VISUALIZAÇÃO COMPUTACIONAL

Construção do *dashboard* Streamlit com Plotly (9 visualizações interativas)

3.4.6 PROJETOS E MÉTODOS PARA A PRODUÇÃO DO CONHECIMENTO

Estruturação do relatório e formatação ABNT

3.4.7 ENGENHARIA DE SOFTWARE

Versionamento com Git, modularização do código e documentação (README)

4 CONCLUSÃO E RECOMENDAÇÕES

4.1 CONCLUSÃO

O estágio realizado na Musakanon Serviços proporcionou uma experiência completa e desafiadora em Ciência de Dados aplicada. O projeto demonstrou que é viável construir, com ferramentas *Open Source* e execução em CPU, um *pipeline* automatizado capaz de transformar horas de discurso parlamentar em *insights* estruturados e acessíveis.

Os resultados alcançados validam a aplicabilidade de técnicas de NLP ao contexto legislativo brasileiro. O cruzamento entre tópicos e sentimentos revelou padrões relevantes, como a forte carga negativa em debates sobre segurança pública e corrupção, e a predominância de tom positivo em discussões sobre projetos de lei e iniciativas comunitárias.

Do ponto de vista da formação profissional, o estágio consolidou competências em todas as fases do ciclo de dados: coleta, processamento, modelagem, avaliação e comunicação de resultados. A necessidade de fazer escolhas técnicas fundamentadas (como, por exemplo, a seleção do modelo de sentimento), lidar com limitações reais (como o tópico genérico que absorveu 30% dos segmentos) e comunicar resultados de forma clara em um *dashboard* interativo, foram aprendizados diretamente aplicáveis à carreira em Ciência de Dados.

O código-fonte completo do projeto, incluindo todos os *scripts* do *pipeline*, *dashboard* e documentação, encontra-se disponível em repositório Git, permitindo reprodutibilidade e extensão futura por qualquer pesquisador ou desenvolvedor interessado em transparência legislativa.

4.2 SUGESTÕES

Com base na experiência adquirida durante o estágio, apresentam-se as seguintes sugestões de melhoria e evolução para o projeto e para a empresa:

4.2.1 PARA O PROJETO

- Diarização de falantes: incorporar um módulo de *speaker diarization* para identificar automaticamente qual vereador está falando em cada segmento. Ferramentas como o `pyannote.audio` permitiriam cruzar análise de sentimento com posicionamento político individual.

- *Fine-tuning* do modelo de sentimento: treinar o XLMM-RoBERTa com dados anotados de discurso parlamentar brasileiro para melhorar a distinção entre frases procedimentais neutras e expressões de opinião.
- Integração com APIs oficiais: conectar o *pipeline* à API de dados abertos da Câmara Municipal para automatizar completamente a coleta de novas sessões à medida que são publicadas

4.2.2 PARA A EMPRESA

O projeto demonstrou a viabilidade de oferecer soluções de NLP como serviço para organizações da sociedade civil, mídia e governos. A empresa poderia explorar este nicho comercializando *dashboards* analíticos personalizados para diferentes câmaras municipais e assembleias legislativas, com processamento semanal automatizado e baixo custo operacional (apenas CPU necessária).

4.3 RESULTADO

Figura 1 - Dashboard: Métricas Gerais e Distribuição de Sentimento



Figura 2 - Dashboard: *Heatmap*: Tópico x Sentimento

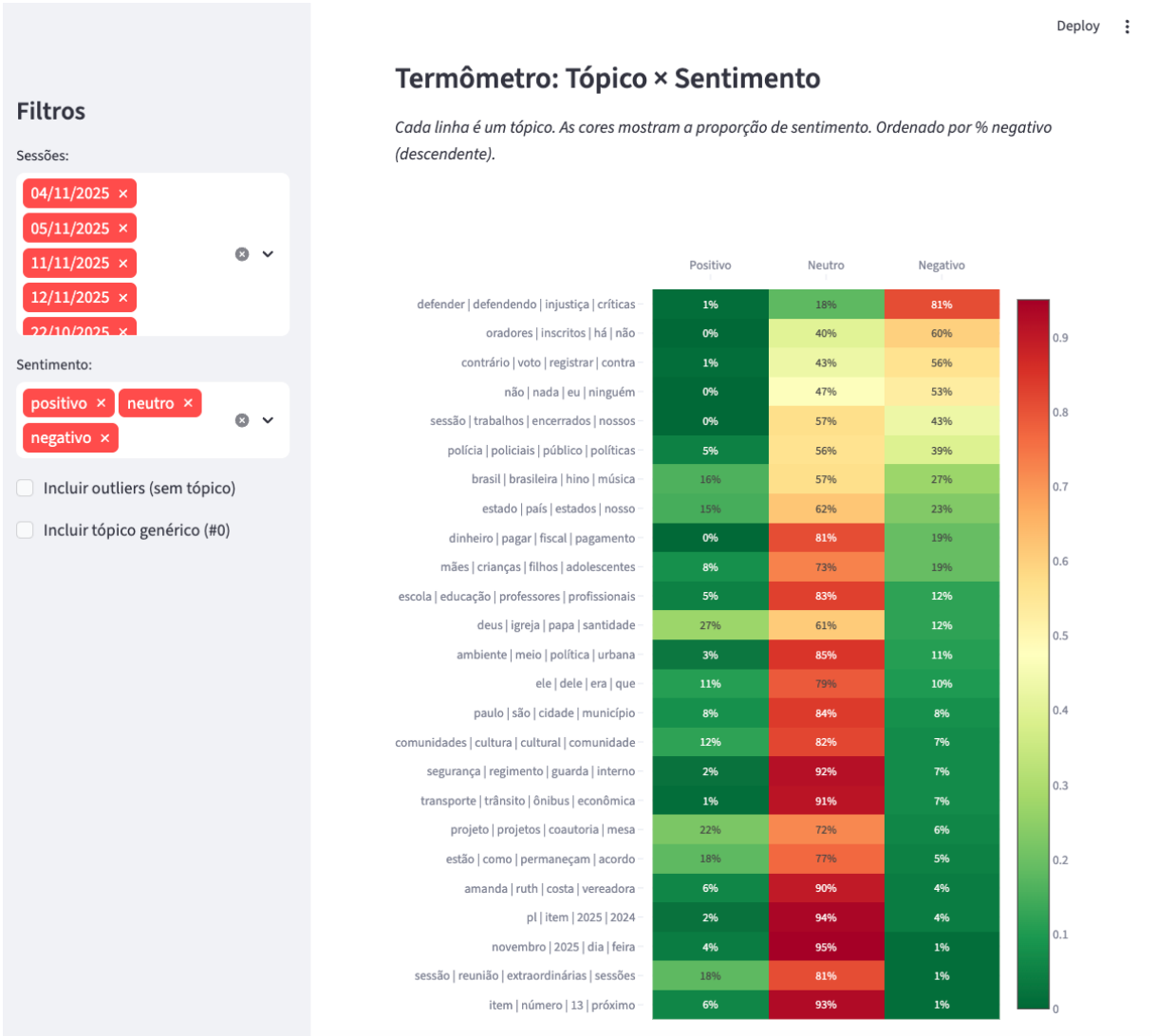


Figura 3 - Dashboard: Top tópicos por volume

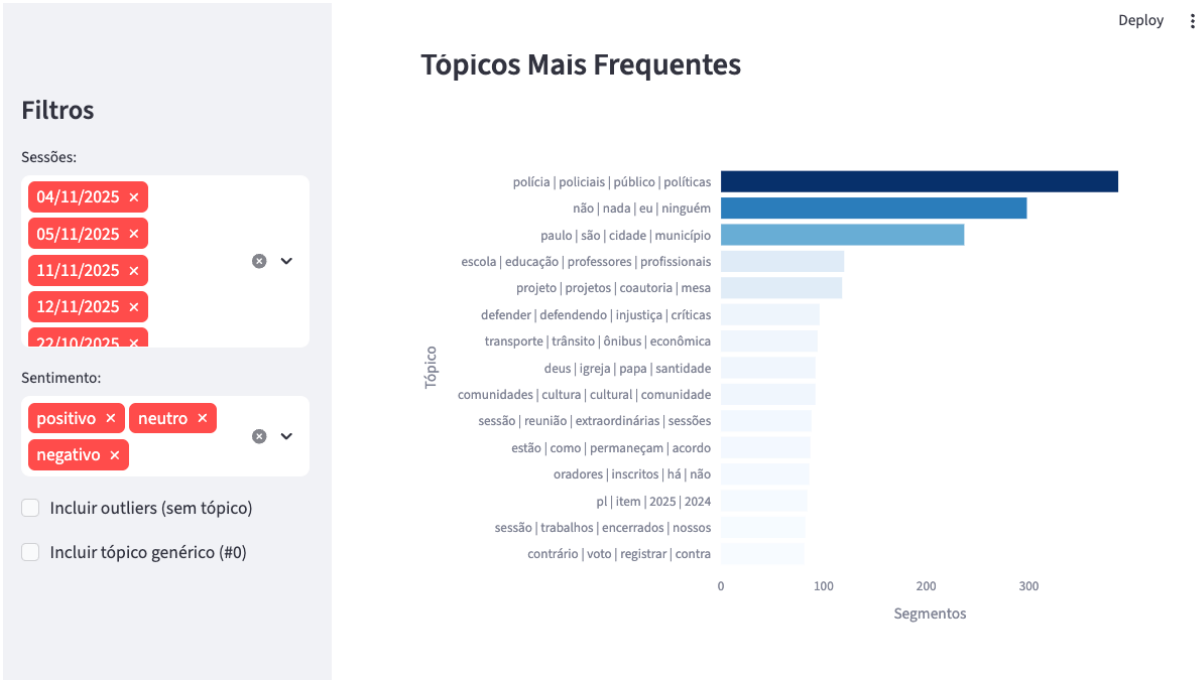


Figura 5 - Dashboard: Distribuição de confiança por sentimento

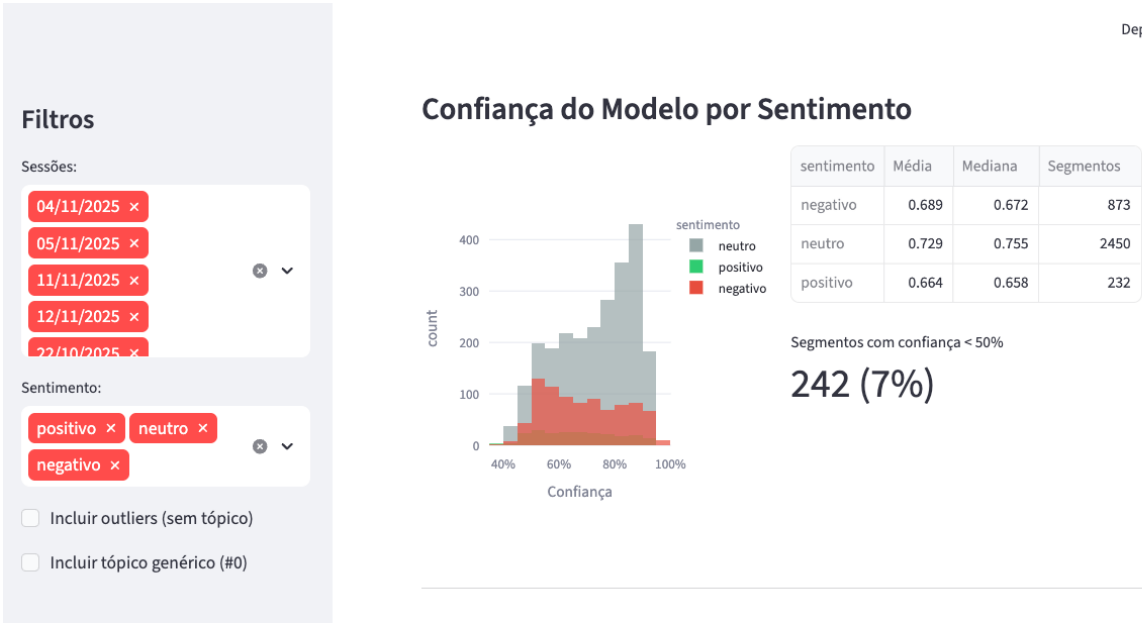


Figura 7 - Dashboard: Nuvem de palavras



Fonte: Elaborado pela autora (2026).

5 REFERÊNCIAS

3Blue1Brown. **Neural Networks**. YouTube, *playlist*. Disponível em: https://youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi. Acesso em: 2 fev. 2026.

BARBIERI, Francesco; ANKE, Luis Espinosa; CAMACHO-COLLADOS, Jose. **XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond**. arXiv, 11 maio 2022. Disponível em: <http://arxiv.org/abs/2104.12250>. Acesso em: 10 fev. 2026

DEVLIN, Jacob et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. In: **Proceedings of NAACL-JTL**, 2019, Minneapolis. Proceedings[...]. Association for Computational Linguistics, 2019. p. 4171-4186.

GROOTENDORST, Maarten. **BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure**. arXiv preprint, arXiv:2203.05794, 2022. Disponível em: <https://arxiv.org/abs/2203.05794>. Acesso em 11 fev. 2026.

HONNIBAL, Matthew; MONTANI, Ines. **spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing**. 2017. Disponível em: <https://spacy.io/>. Acesso em 9 fev. 2026.

RADFORD, Alec *et al.* **Robust Speech Recognition via Large-Scale Weak Supervision**. arXiv, 6 dez. 2022. Disponível em: <http://arxiv.org/abs/2212.04356>. Acesso em: 6 fev. 2026

UNIVESP. **Plano de Ensino: COM300 - Banco de Dados**. UNIVESP. Disponível em: <https://assets.univesp.br/blackboard/plano-de-ensino/disciplinas/COM300.html>. Acesso em 27 fev. de 2026.

UNIVESP. **Plano de Ensino: COM390 - Engenharia de Software**. UNIVESP. Disponível em: <https://assets.univesp.br/blackboard/plano-de-ensino/disciplinas/COM390.html>. Acesso em 27 fev. de 2026.

UNIVESP. **Plano de Ensino: COM400 - Visualização Computacional**. UNIVESP. Disponível em: <https://assets.univesp.br/blackboard/plano-de-ensino/disciplinas/COM400.html>. Acesso em 27 fev. de 2026.

UNIVESP. **Plano de Ensino: COM410 - Aprendizado de Máquina.** UNIVESP. Disponível em: <https://assets.univesp.br/blackboard/plano-de-ensino/disciplinas/COM410.html>. Acesso em 27 fev. de 2026.

UNIVESP. **Plano de Ensino: COM550 - Processamento de Linguagem Natural.** UNIVESP. Disponível em: <https://assets.univesp.br/blackboard/plano-de-ensino/disciplinas/COM550.html>. Acesso em 27 fev. de 2026.

UNIVESP. **Plano de Ensino: INT100 - Projetos e Métodos para a Produção do Conhecimento.** UNIVESP. Disponível em: <https://assets.univesp.br/blackboard/plano-de-ensino/disciplinas/INT100.html>. Acesso em 27 fev. de 2026.

UNIVESP. **Plano de Ensino: PES300 - Estatística e Probabilidade.** UNIVESP. Disponível em: <https://assets.univesp.br/blackboard/plano-de-ensino/disciplinas/PES300.html>. Acesso em 27 fev. de 2026.

ANEXO - FICHA DE PRESENÇA



ESTÁGIO SUPERVISIONADO
Bacharelado em Ciência de Dados
FICHA DE PRESENÇA

ESTAGIÁRIO

Nome completo:	Cintia Izumi Shinoda
RA:	
Curso:	Bacharelado em Ciência de Dados
Polo:	Polo UAB CEU - São Paulo - Parque Bristol

PERÍODO DO ESTÁGIO

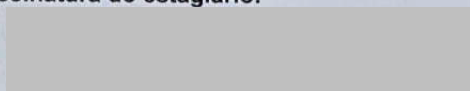
Data de início do estágio:	Data de finalização do estágio:
02/02/2026	27/02/2026

EMPRESA/INSTITUIÇÃO CONCEDENTE DO ESTÁGIO

Nome da concedente: MUSAKANON SERVICOS - SOCIEDADE UNIPessoal LTDA	
Endereço	
Número:	Complemento:
Bairro:	CEP:
Cidade: São Paulo	Estado: São Paulo
Contato	E-mail:
	Telefone:

SUPERVISOR LOCAL DE ESTÁGIO

Nome do supervisor:	Ademir Stefanski Junior
Assinatura:	Rubrica:

Assinatura do estagiário:**Nome do
estagiário:**

Cintia Izumi Shinoda

RELATÓRIO DE ATIVIDADES

	Data	Nº de horas	Atividade desenvolvida	Rubrica do Supervisor Local
1	02/02/2026	3,5	Estudo de conceitos fundamentais de <i>Natural Language Processing</i> (NLP) e redes neurais.	
2	02/02/2026	1	Entender o <i>pipeline</i> do projeto.	
3	02/02/2026	0,5	Instalação de ferramentas (Python, Git, VS Code).	
4	02/02/2026	0,5	Criação do Ambiente Virtual e Instalação e teste das dependências	
5	02/02/2026	0,5	Criação da Estrutura de Pastas	
6	03/02/2026	2	Estudo de Speech-to-Text e Whisper	
7	03/02/2026	3	Estudo sobre Transformers e BERT	
8	03/02/2026	1	Mapeamento das fontes de dados (YouTube)	
9	04/02/2026	2	Estudo do yt-dlp	
10	04/02/2026	0,5	Identificar URLs de 6 vídeos para o MVP	
11	04/02/2026	2	Escrever código para fazer download dos áudios (<code>download_audio.py</code>)	
12	04/02/2026	1,5	Fazer o download dos vídeos	
13	05/02/2026	2	Criar catálogo de metadados no DuckDB	
14	05/02/2026	0,5	Consultar catálogo via SQL	
15	05/02/2026	0,5	Validar integridade dos áudios	
16	05/02/2026	2	Documentação	
17	05/02/2026	1	Reunião com supervisor	
18	06/02/2026	2	Escolher modelo e estudar parâmetros	
19	06/02/2026	3	Escrever código para transcrever os áudios (<code>transcribe.py</code>)	
20	06/02/2026	1	Validar transcrições	
21	09/02/2026	0,5	Atualizar DuckDB para confirmar status	

22	09/02/2026	1	Analisar pontos com problemas (vinhetas, segmentos curtos, redundância e texto sem valor analítico)
23	09/02/2026	2,5	Escrever código para realizar pré-processamento (<code>preprocess.py</code>)
24	09/02/2026	2	Rodar <i>batch</i> completo
25	10/02/2026	0,5	Atualização no DuckDB
26	10/02/2026	1,5	Escolher Modelo de Sentimento (<code>cardiffnlp</code>)
27	10/02/2026	2	Escrever código do <code>sentiment.py</code>
28	10/02/2026	1	Rodar <i>batch</i> completo
29	10/02/2026	1	Reunião com supervisor
30	11/02/2026	2	Estudo do funcionamento do BERTopic
31	11/02/2026	3	Escrever o código para extração dos tópicos (<code>topics.py</code>)
32	11/02/2026	1	Fazer ajuste e rodar novamente
33	12/02/2026	1,5	Planejar o Dashboard
34	12/02/2026	4,5	Escrever o código para criar o <i>Dashboard</i> (<code>dashboard.py</code>)
35	13/02/2026	1,5	Refinamento visual
36	13/02/2026	3	Preparar material para reunião
37	13/02/2026	1,5	Atualização do README
38	19/02/2026	2	Validação qualitativa
39	19/02/2026	2,5	Reduzir o tópico #0
40	19/02/2026	1,5	Adicionar visualização temporal por tópico
41	20/02/2026	2	Adição da visualização de Confiança do modelo
42	20/02/2026	2	Reunião com supervisor
43	20/02/2026	2	Escrever código <code>run_pipeline.py</code>
44	23/02/2026	3	Teste <i>end-to-end</i> do <i>pipeline</i>
45	23/02/2026	3	Observação e avaliação crítica do <i>dashboard</i> .
46	24/02/2026	2	Reunião com supervisor
47	24/02/2026	4	Documentação das observações.

48	25/02/2026	2	Revisão do <i>pipeline</i> completa.	
49	25/02/2026	4	Documentação das observações	
50	26/02/2026	3	Revisão do <i>pipeline</i> e documentação das observações	
51	26/02/2026	2	Reunião com supervisor	
52	26/02/2026	0,5	Verificação dos docstrings	
53	26/02/2026	0,5	Atualização das dependências (requirements.txt)	
54	27/02/2026	4	Redação do Relatório Final	

**TOTAL DA CARGA HORÁRIA
CUMPRIDA**

100 horas