

Prémio Arquivo.pt

Descrição Sumária do Trabalho (documento-modelo)

Instruções gerais:

- Apresente de forma concisa os contributos do trabalho de acordo com os critérios de avaliação.
- Responda diretamente às questões colocadas.
- Use linguagem clara e sucinta.
- Pode incluir imagens e respetivas legendas.
- O tamanho máximo do ficheiro é de 10 MB.
- O limite de palavras é de 2 000.
- Submeta a versão final num dos seguintes formatos: DOCX ou PDF.

Identificação

- Título: **Narrative Monitoring**: Análise de Teorias Conspirativas de Substituição Populacional no Arquivo Web Português (1996-2021)
- Área temática: Análise de discursos digitais, História dos media online, Sociologia digital
- Candidato: [Erik Bran Marino](#), Rafael Prezado, Ana Sofia Ribeiro, Renata Vieira
- Email: erik.marino@uevora.pt m57355@alunos.uevora.pt asvribeiro@uevora.pt renatav@uevora.pt

Descrição do Trabalho

O trabalho "Narrative Monitoring" constitui uma análise sistemática e abrangente da emergência e evolução das teorias conspirativas de substituição populacional (PRCT), no espaço digital português, entre 1996 e 2021. Utilizando o Arquivo.pt como fonte primária, desenvolvemos uma metodologia que combina técnicas de *web archiving*, processamento de linguagem natural e análise estatística para identificar, classificar e analisar 36.621 documentos relacionados com temáticas migratórias e demográficas.

A plataforma web desenvolvida (<https://www.narrativemonitor.uevora.pt/>) apresenta os resultados através de visualizações interativas, uma linha cronológica dinâmica e um quiz educativo sobre as palavras-chave mais associadas a discursos conspiratórios. O objetivo é proporcionar aos investigadores e ao público uma ferramenta para compreender a evolução destes discursos em Portugal e promover a literacia mediática.

Objetivos

1. **Criar um corpus abrangente** de conteúdos web históricos relacionados com migração, demografia e teorias conspirativas a partir do Arquivo.pt (1996-2021).
2. **Desenvolver e aplicar uma metodologia** para classificação automática de documentos em "PRCT" (conteúdo conspiratório) e "Non-PRCT".
3. **Analisar padrões temporais** da presença e frequência destas narrativas no panorama digital português.
4. **Identificar fontes principais** de disseminação destas teorias e a sua concentração.
5. **Disponibilizar ferramentas interativas** para educação e consciencialização do público sobre discursos problemáticos.
6. **Contribuir para a compreensão** da evolução dos discursos extremistas online em Portugal.

Resultados Atingidos

Os principais resultados obtidos através da análise do corpus de 36.621 documentos foram:

1. **Distribuição temporal em números absolutos:**
 - 1996-1998: 0 documentos PRCT de um total de 5.411 documentos
 - 1999-2009: 30 documentos PRCT de um total de 8.493 documentos
 - 2010-2016: 226 documentos PRCT de um total de 7.723 documentos
 - 2017: 84 documentos PRCT de um total de 288 documentos
 - 2018: 112 documentos PRCT de um total de 3.720 documentos
 - 2019: 338 documentos PRCT de um total de 3.547 documentos
 - 2020: 47 documentos PRCT de um total de 3.275 documentos
 - 2021: 1 documento PRCT de um total de 278 documentos
2. **Distribuição temporal em percentagens:**
 - 1996-2009: Quase ausência (média 0,51%)
 - 2010-2017: Crescimento acentuado (média 6,21%)
 - 2017: Pico histórico (29,17%)
 - 2018-2021: Declínio gradual (média 3,58%)
3. **Concentração em fontes específicas:**
 - Blogs: 9,78% dos conteúdos são PRCT (representando a maioria dos conteúdos conspiratórios)
 - Media tradicionais: Apenas 0,21% dos conteúdos classificados como PRCT

- 18,02% de todos os conteúdos PRCT (aproximadamente 151 documentos) provêm de um único domínio (gladio.blogspot.com)
- 4. **Identificação de palavras-chave significativas:**
 - "Islamização da Europa", "elite globalista" e "agenda globalista" apresentam percentagens de PRCT acima de 95%
 - "Plano Kalergi" e "agenda Kalergi" com percentagens acima de 84%
- 5. **Plataforma interativa funcional** integrando visualizações dinâmicas, linha cronológica e quiz educativo
- 6. **Dataset estruturado e metodologia replicável** para futuras investigações, disponível mediante solicitação para fins de investigação académica, em conformidade com as políticas da plataforma e as normativas nacionais e europeias de proteção de dados.
- 7. **Processo de Classificação PRCT/Non-PRCT:** Para a classificação automática dos documentos em PRCT e Non-PRCT, implementámos uma metodologia híbrida utilizando a API DeepSeek. Definimos claramente os critérios operacionais para cada categoria, onde documentos contendo menções a teorias como "Grande Substituição", "Plano Kalergi" ou representações da imigração como "guerra demográfica" foram classificados como PRCT. A validação manual de uma amostra balanceada de 40 documentos por dois anotadores independentes revelou:
 - PABAK (Kappa Ajustado para Prevalência e Viés): 0,8000, indicando um acordo substancial entre anotadores em dados desbalanceados
 - Precisão global da validação: 90,00% de concordância entre os dois anotadores
 - Taxa de documentos classificados com sucesso: 88,56% do conjunto inicial

Originalidade e carácter inovador

O trabalho apresenta diversos elementos inovadores:

1. **Primeira análise quantitativa sistemática** sobre teorias conspirativas de substituição populacional no contexto português, preenchendo uma lacuna significativa na literatura académica.
2. **Metodologia híbrida** que combina:
 - Extração e processamento de dados históricos do Arquivo.pt,
 - Classificação automática através de APIs de IA,
 - Análise estatística rigorosa,
 - Visualização interativa de dados temporais.
3. **Abordagem diacrónica abrangente** (25 anos) que permite identificar padrões de médio e longo prazo e relacionar com eventos sócio-políticos.
4. **Sistema de classificação de fontes** desenvolvido especificamente para este projeto, uma vez que o Arquivo.pt não fornece categorização a priori dos domínios. A nossa equipa desenvolveu uma taxonomia própria que categoriza os domínios em 8 tipos

principais (Blog, Social Media, Other, News/Media, Government, Academic, Encyclopedia, Forum) com base em características estruturais e de conteúdo. Por exemplo, domínios com extensões ".gov.pt" ou contendo termos como "governo", "ministério", "câmara" foram classificados como "Government"; domínios com termos como "universidade", "faculdade", "instituto" e publicações académicas foram classificados como "Academic"; enquanto plataformas como "blogspot.com" e "wordpress.com" foram classificadas como "Blog". Esta categorização permitiu análises diferenciadas por tipo de média e uma melhor compreensão das fontes de disseminação das narrativas PRCT.

5. **Componente educativa interativa** através do quiz "Conspiracional ou não?" que transforma a investigação numa ferramenta pedagógica.

Impacto social (aplicação e utilidade social)

O trabalho oferece diversas contribuições sociais:

1. **Promoção da literacia mediática:** A plataforma permite aos cidadãos identificar e compreender narrativas conspiratórias, funcionando como ferramenta educativa para aumentar o pensamento crítico e do próprio conhecimento das PRCT.
2. **Consciencialização pública:** Oferece dados concretos sobre a prevalência e evolução de discursos extremistas online, contribuindo para desmistificar perceções sobre a sua dimensão e para distinguir entre informação verdadeira e conspirativa.
3. **Apoio a educadores e jornalistas:** Fornece recursos valiosos para profissionais que trabalham em áreas de fact-checking, educação para os media e análise de desinformação.
4. **Ferramenta para decisores políticos:** Disponibiliza evidências empíricas sobre padrões de disseminação de narrativas problemáticas para informar políticas públicas.
5. **Preservação da memória coletiva digital:** Valoriza o património digital português, demonstrando a importância de preservar conteúdos web para análise sociológica e histórica.

Impacto científico (aplicação e utilidade científica)

Para a comunidade científica, o trabalho oferece:

1. **Metodologia replicável:** O pipeline de extração, classificação e análise pode ser adaptado para estudar outros tipos de discursos problemáticos em arquivos web.
2. **Dataset estruturado:** Disponibiliza um corpus anotado sinteticamente de 36.621 documentos que pode servir de base para futuras investigações comparativas.
3. **Análise longitudinal inédita:** Fornece dados quantitativos sobre a evolução temporal de narrativas conspiratórias, permitindo correlações com eventos sociopolíticos.
4. **Framework para análise de fontes:** O sistema de categorização de domínios permite estudos comparativos entre diferentes tipos de media.
5. **Caso de estudo português:** Contribui para estudos comparativos internacionais, permitindo analisar como estas narrativas evoluem em diferentes contextos nacionais ou, mesmo transnacionais e transculturais.
6. **Ponte entre disciplinas:** Integra métodos de ciência de dados, análise de discurso, sociologia e história dos media, numa abordagem interdisciplinar.

Relevância da utilização do Arquivo.pt

O Arquivo.pt foi absolutamente essencial para este trabalho pelas seguintes razões:

1. **Acesso a conteúdos históricos extintos:** Grande parte dos blogs e websites analisados já não estão ativos, sendo o Arquivo.pt a única forma de aceder a estes conteúdos. Por exemplo, domínios como "omarxismocultural.blogspot.pt" (4,53% dos PRCT) e "totalitarismouniversalista.blogspot.com" (2,86% dos PRCT) só podem ser consultados através do arquivo.
Exemplo concreto: Em 2017, quando se verificou o pico percentual de narrativas PRCT (29,17%), sites como "gladio.blogspot.com" publicaram extensivamente sobre o "Plano Kalergi". Sem o Arquivo.pt, estes conteúdos teriam sido perdidos, impossibilitando a compreensão deste momento crucial na disseminação destas teorias em Portugal.
2. **Cobertura temporal abrangente:** A preservação de 25 anos (1996-2021) de conteúdos web permitiu a análise diacrónica completa do fenómeno PRCT em Portugal, revelando a sua emergência, o crescimento e o decréscimo.
3. **API completa e robusta:** A API do Arquivo.pt possibilitou a extração sistemática de conteúdos usando aproximadamente 150 palavras-chave portuguesas relacionadas com migração e mudanças demográficas, resultando num corpus abrangente. O processo de seleção das 150 palavras-chave baseou-se numa abordagem abrangente

do campo semântico da migração e das mudanças demográficas. Foram consideradas tanto palavras neutras de uso comum em discursos migratórios (ex: "imigração", "refugiados", "demografia") quanto termos específicos associados a narrativas conspirativas (ex: "Plano Kalergi", "substituição populacional"). Para garantir a máxima cobertura, incluíram-se raízes lexicais, variantes morfológicas e diferentes formulações dos mesmos conceitos, permitindo capturar as diversas formas como estes temas foram discutidos ao longo do período analisado.

4. **Preservação de metadados cruciais:** O Arquivo.pt mantém informações essenciais como URLs originais, datas de captura e estrutura dos sites, fundamentais para a análise contextual dos conteúdos.

Comentários adicionais

Este trabalho demonstra o valor do Arquivo.pt não apenas como repositório histórico, mas como ferramenta de investigação para compreender fenómenos sociais contemporâneos. A metodologia desenvolvida pode ser aplicada a outros contextos linguísticos e temáticos, contribuindo para uma compreensão mais ampla da evolução dos discursos online.

O desenvolvimento da plataforma interativa representa um esforço para tornar a investigação académica acessível ao público geral, transformando dados complexos em visualizações compreensíveis e ferramentas educativas.

O desenvolvimento técnico da plataforma integra um BackEnd desenvolvido com Django e um FrontEnd construído com React Framework, otimizado com Vite JS. No BackEnd, foram criados modelos específicos para cada ano entre 1996 e 2021, refletindo a estrutura dos dados extraídos, com campos como `document_id`, `classification`, `keyword`, `title`, `URLs` e `text_content`. O deploy foi realizado com Docker e Docker Compose, utilizando Gunicorn como servidor WSGI e PostgreSQL como base de dados.

Recursos complementares

- Narrative Monitoring Platform, <https://www.narrativemonitor.uevora.pt/>: Plataforma interativa desenvolvida no âmbito deste projeto.