Prémio Arquivo.pt

Descrição Sumária do Trabalho

Identificação

- **Título:** Lupa Digital
- **Área temática:** Ciência de Dados, Metajornalismo, Literacia Mediática
- **Candidato:** Hugo Veríssimo
- **Email:** hugoverissimo21@gmail.com

Descrição do Trabalho

A Lupa Digital (link) é um projeto desenvolvido no âmbito da competição Prémios Arquivo.pt 2025, focado em metajornalismo inteligente, com o objetivo de combater a parcialidade e tornar a informação acessível e transparente a qualquer pessoa.

Este projeto propõe transformar grandes volumes de informação jornalística arquivada em conhecimento útil e acessível. Com base nos milhões de páginas do Arquivo.pt e recorrendo a ferramentas inovadoras, como modelos de processamento de linguagem natural (NLP), a Lupa Digital torna possível a análise e exploração de entidades, eventos e temas ao longo do tempo, contribuindo para o reforço da literacia mediática.

Com notícias arquivadas desde 1998 até à atualidade, de 20 fontes de informação nacionais, a Lupa Digital, para além do desenvolvimento de um conjunto de dados com centenas de milhares de notícias e metadados referentes à sua fonte, data de publicação, perceção do sentimento e tópicos mencionados, consiste também numa aplicação *web* que permite ao utilizador explorar qualquer tópico à sua escolha.

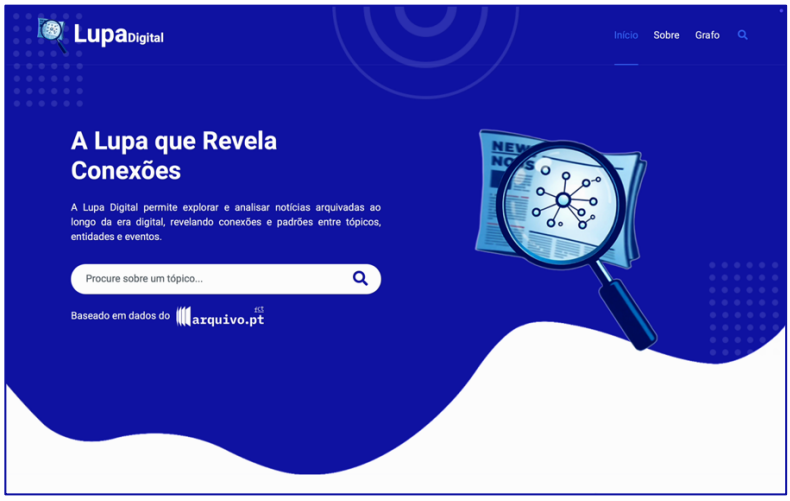

Figura 1. Captura de ecrã da página principal da Lupa Digital.

Esta aplicação *web*, após a pesquisa por parte do utilizador de um tópico do seu interesse, gera várias visualizações interativas sobre as aparições do mesmo nos media digitais portugueses, respondendo a perguntas tais como:

- Que fontes têm noticiado mais esse determinado tópico?

- Como têm evoluído as suas menções ao longo dos anos?

- Que outros tópicos se relacionam com este e como têm variado ao longo dos anos?

- Quais são as principais relações deste tópico e qual é a sua perceção?

Objetivos

Os principais objetivos da Lupa Digital incluem:

- Análise da abordagem dos media em relação a qualquer tópico

Permitir que qualquer utilizador, de forma simples e rápida, consiga pesquisar um tópico à sua escolha e, a partir dos resultados dessa pesquisa, consiga ter uma perceção em relação a como os media portugueses tendem a abordar esse tema e como é que esta abordagem tem vindo a mudar ao longo dos anos.

- Identificação e análise de relações entre tópicos

Revelar ligações entre qualquer par de tópicos, sejam eles entidades, eventos ou temas, e, por conseguinte, permitir ao utilizador ter uma perceção sobre a relação, no que toca a verificar quais as fontes de informação mais envolvidas na ponte entre os tópicos, o sentimento transmitido no contexto das notícias onde a relação está presente e como tem variado a ligação ao longo do tempo.

- Promoção de transparência e literacia mediática

Incentivar a análise crítica da informação, ajudando os utilizadores a compreender a evolução dos discursos e a influência dos media na opinião pública, de modo que os mesmos possam ter uma perceção imparcial através dos resultados apresentados pela Lupa Digital.

- Criação de uma ferramenta interativa e intuitiva

Desenvolver uma plataforma inovadora e acessível a qualquer pessoa, com uma interface simples e intuitiva, onde os utilizadores possam explorar dados relacionados com os media portugueses de forma visual e interativa.

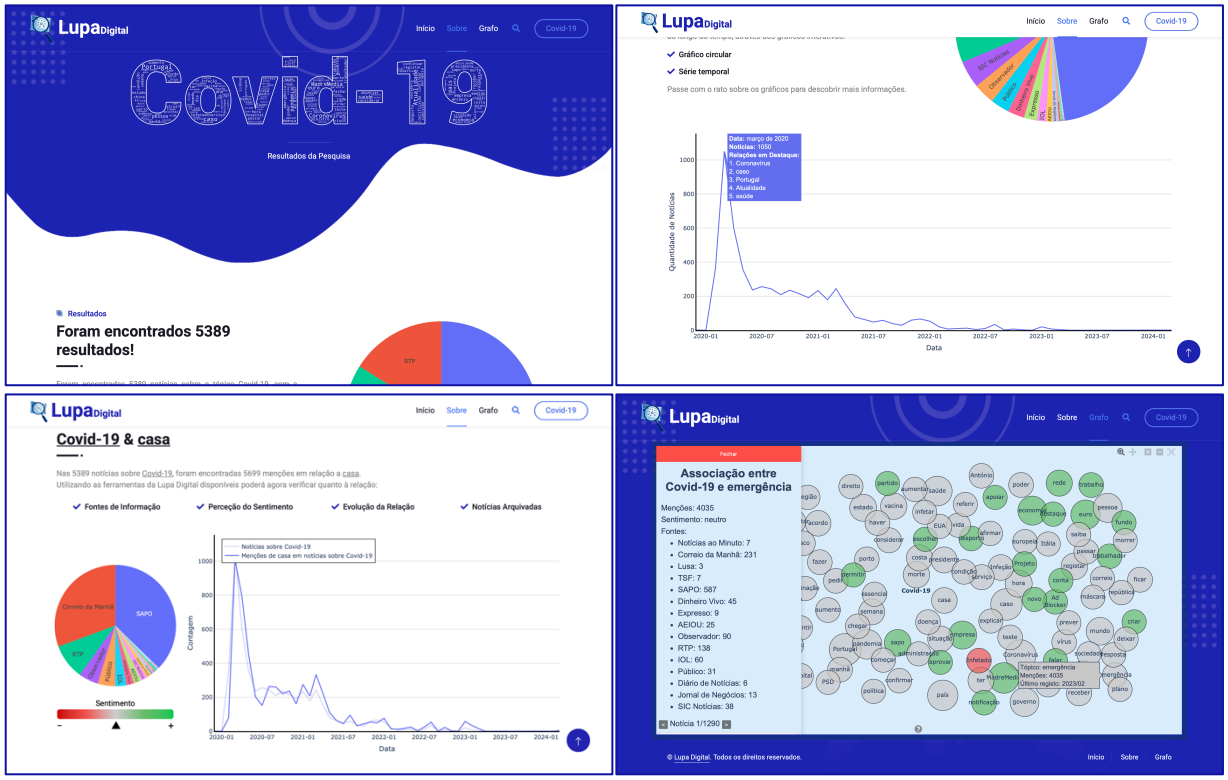

Figura 2. Exemplos da interface da plataforma Lupa Digital, ilustrando as principais funcionalidades: pesquisa por tópicos (ex.: "Covid-19"), visualização da evolução temporal da cobertura mediática e dos sentimentos associados, bem como a identificação de relações entre temas através de grafos interativos. Estas ferramentas permitem uma análise crítica e intuitiva da informação mediática, promovendo transparência, literacia mediática e exploração de dados de forma acessível.

Resultados Atingidos

Através da criação da Lupa Digital foram obtidos dois resultados principais: um conjunto de dados, disponível em acesso aberto, e uma aplicação *web*.

Atendendo ao conjunto de dados (disponível em doi.org/10.5281/zenodo.15231163), como necessidade para o desenvolvimento do trabalho, este foi criado a partir do tratamento de dados extraídos do Arquivo.pt. É um conjunto de dados com mais de três centenas de milhares de notícias, desde 1998 até à atualidade, de 20 fontes de informação portuguesas, nomeadamente AEIOU, Correio da Manhã, CNN Portugal, Diário de Notícias, Dinheiro Vivo, Expresso, IOL, Jornal de Notícias, Lusa, Jornal de Negócios, NiT, Notícias ao Minuto, O Mirante, Observador, Público, Record, RTP, Sapo, SIC Notícias, e TSF, contendo metadados sobre cada uma delas, tais como a perceção do sentimento transmitido pela notícia e os tópicos mencionados na mesma. Este conjunto de dados para além de permitir futuros estudos, também permite a sua expansão futura com notícias e artigos mais recentes e com a adição de novas fontes de informação.

Quanto à aplicação *web* (link), sendo esta o foco principal da Lupa Digital, envolveu a criação de website que funciona como um motor de busca para qualquer tópico que um utilizador tenha interesse. Após a pesquisa por um tópico, o utilizador terá acesso a várias visualizações e informações relacionadas com a pesquisa, podendo, de forma intuitiva, analisar criticamente as relações com outros tópicos, como é que a perceção do tópico pesquisado tem vindo a mudar e as divergências entre as menções de diferentes fontes de informação no que toca ao seu tópico.

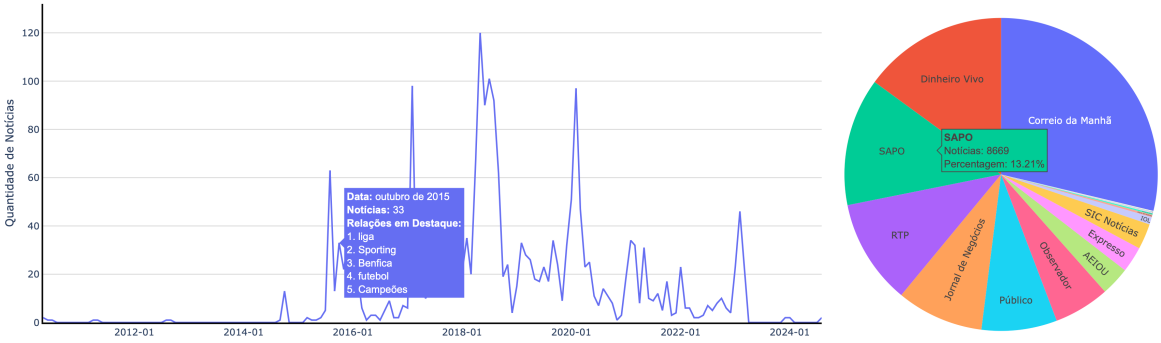

Figura 3. Exemplo de duas visualizações interativas presentes na aplicação *web*. Uma série temporal a representar a quantidade de notícias sobre o tópico pesquisado e relações em destaque e um gráfico circular a revelar a distribuição das notícias sobre o tópico pesquisado entre as várias fontes de informação.

Originalidade e caráter inovador

A Lupa Digital distingue-se pela sua abordagem inovadora no que toca ao metajornalismo, pelo facto de usar técnicas de automatização para o tratamento de texto e processamento de linguagem natural, em larga escala e de forma eficiente, permitindo uma exploração de uma maior quantidade de arquivos digitais e, desta forma, uma experiência interativa e dinâmica na análise de milhares de notícias preservadas no Arquivo.pt.

Para além disso, ao invés de uma simples pesquisa textual, o projeto proporciona várias visualizações gráficas sobre os tópicos pesquisados e as relações com os demais, permitindo que os utilizadores identifiquem conexões e padrões de forma intuitiva e transparente. Estas visualizações gráficas incluem um grafo interativo que revela como diferentes tópicos se interligam ao tópico pesquisado, e ao interagir com cada tópico permite que o utilizador obtenha variada informação sobre a relação em causa, tal como a quantidade de notícias que contam com a presença dessa relação, as fontes de notícias que mais a mencionam, a perceção do sentimento associada à mesma, e como tem vindo a evoluir a quantidade de menções da relação.

Este projeto não é apenas um motor de busca para notícias, mas também uma ferramenta de investigação e análise que ajuda a compreender como os media moldaram a perceção pública ao longo do tempo. A possibilidade de acompanhar a evolução de narrativas e identificar padrões entre fontes de notícias ou entre relações de tópicos tornam este projeto um recurso único para a literacia mediática.

Outro contributo inovador é a criação, e partilha através de acesso aberto, de um dos maiores, tanto em quantidade como em horizonte temporal, conjunto de dados no âmbito da mineração de dados em notícias, e neste caso particular, notícias portuguesas.

Impacto social (aplicação e utilidade social)

O projeto Lupa Digital tem uma grande utilidade para os cidadãos, pelo facto de disponibilizar uma ferramenta inovadora que permite aos mesmos explorar e compreender a evolução de um tópico do seu interesse no contexto da informação jornalística com o passar dos anos, de forma rápida e intuitiva para o público em geral.

Ademais, num contexto em que o acesso à informação e a literacia mediática são essenciais no dia-a-dia dos portugueses, a Lupa Digital incentiva a compreensão mais profunda da informação e a análise crítica das fontes noticiosas, através da democratização do acesso a dados históricos a partir de uma simples pesquisa.

Atendendo à funcionalidade do grafo interativo, este ajuda a visualizar as conexões entre temas, entidades e eventos, permitindo aos utilizadores explorar padrões e relações que, de outra forma, poderiam passar despercebidos.

Por fim, a acessibilidade facilitada a uma grande quantidade de informação, para além de contribuir para o combate à desinformação, beneficia desde estudantes até cidadãos comuns interessados em várias áreas do saber (economia, história, política, ...), promovendo um maior envolvimento com a informação digital preservada.

Impacto científico (aplicação e utilidade científica)

Este projeto torna-se extensamente útil para investigadores e para a comunidade científica em geral, pelo facto de ter desenvolvido, para além de uma ferramenta inovadora para a análise de grandes volumes de notícias arquivadas, um conjunto de dados com cerca de 350 mil notícias e informações associadas às mesmas, tal como a perceção do sentimento e os tópicos mencionados.

O desenvolvimento deste conjunto de dados deixa na mesa inúmeras possibilidades para estudos futuros, tais como estudos sobre temas específicos de forma mais aprofundada, sobre as diferentes fontes de notícias e as suas tendências de comunicação sobre determinados temas, como é que eventos históricos alteraram as narrativas dos meios de comunicação, entre outros. Pela unicidade deste tipo de dados, a sua partilha torna-se assim uma contribuição para o campo do jornalismo digital e investigadores de áreas envolventes.

Ademais, para investigadores da área de Processamento de Linguagem Natural (NLP) e Inteligência Artificial, a Lupa Digital também oferece um caso de estudo valioso, uma vez que combina técnicas avançadas de extração de entidades e análise semântica para estruturar

informação proveniente de diversas fontes e em larga escala. Estes métodos podem ser aplicados e aprimorados em pesquisas relacionadas com desinformação, viés mediático e análise de narrativas.

Relevância da utilização do Arquivo.pt

A partir do uso da CDX-server API do Arquivo.pt foram extraídos mais de três milhões de URLs associados a diversas fontes de informação. Posteriormente, para cada um destes URLs, o Arquivo.pt foi utilizado de modo a permitir a extração do conteúdo associado ao URL, de forma a criar o conjunto de dados que serve de input à aplicação *web* criada. O esquema presente na Fig. 4 demonstra o processo descrito de uma forma mais detalhada.

Este processo revela a importância do Arquivo.pt no desenvolvimento deste projeto, sendo como uma fundação para a criação do conjunto de dados criado e por sua vez para a aplicação *web*. Sem o uso deste arquivo *web* teria sido impossível realizar uma análise da cobertura mediática com um horizontal temporal tão grande (mais de 25 anos), dado ser uma das únicas plataformas que permite o acesso a conteúdos mediáticos mesmo depois de descontinuados.

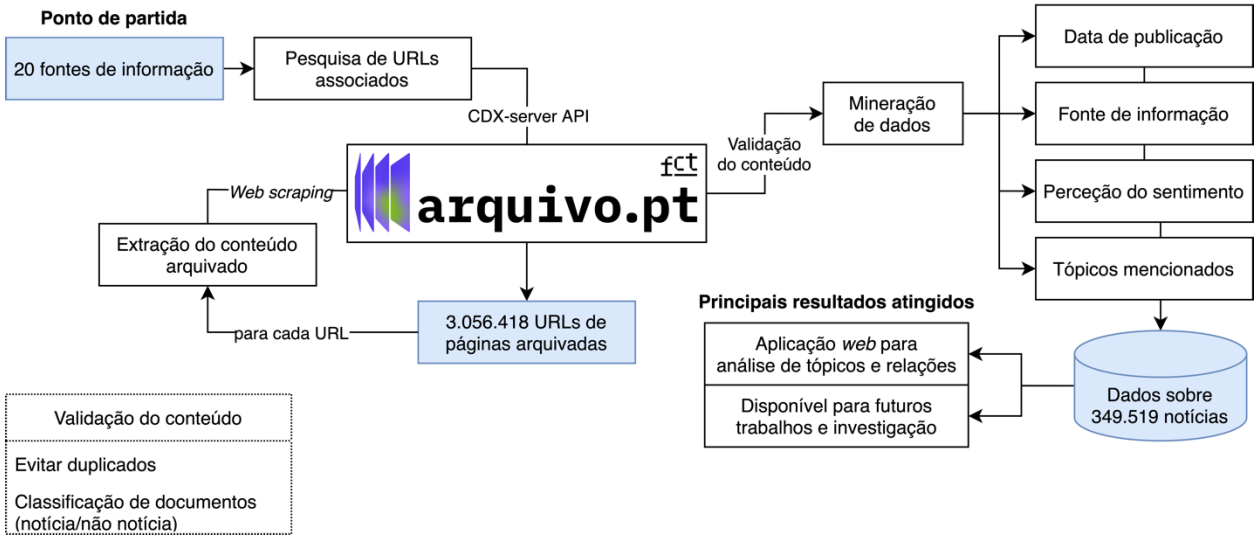

Figura 4. Processo de extração e análise de dados de media digital arquivada.

Comentários adicionais

Este trabalho foi inicialmente desenvolvido em contexto letivo, no âmbito da unidade curricular de Fundamentos de Ciência de Dados (FCD), tendo sido posteriormente aprofundado e melhorado com o objetivo de aplicar conceitos avançados e explorar novas abordagens analíticas. A exploração do Arquivo.pt permitiu não só o desenvolvimento de competências

técnicas em áreas como mineração de dados, processamento de linguagem natural e visualização de informação, como também revelou o potencial desta ferramenta enquanto recurso acessível e útil para quem pretende dar os primeiros passos na área da Ciência de Dados.

Importa também agradecer à equipa do Arquivo.pt pela disponibilização de uma API de acesso gratuito, com documentação clara e bem estruturada, o que facilitou todo o processo de recolha de dados.

Recursos complementares

- Organização do projeto, https://github.com/LupaDigital25

Contém todo o código-fonte desenvolvido no âmbito do projeto, incluindo algoritmos de processamento, análise de dados e interface da aplicação web.

- Aplicação *web*: LINK PARA APP

Página principal da aplicação web desenvolvida.

- *Dataset* criado no âmbito do projeto, https://doi.org/10.5281/zenodo.15231163

Conjunto de dados gerado a partir da análise de todos os URLs associados às fontes de informação analisadas. Inclui metadados, sentimentos extraídos, tópicos e outras anotações relevantes.

- Fonte dos dados do projeto, https://arquivo.pt

Arquivo da *web* portuguesa, utilizado como principal fonte de dados para o projeto.

- API do Arquivo.pt, https://github.com/arquivo/pwa-technologies/wiki/URL-search:-CDX-server-API

Documentação da API do Arquivo.pt, utilizada para a interface com a plataforma do Arquivo.pt.

- Dicionários Natura, https://natura.di.uminho.pt/download/sources/Dictionaries/wordlists/

Coleção de listas de palavras em português, útil para tarefas de processamento de linguagem natural (PLN), como filtragem de *stopwords*, lematização, entre outras. A lista usada foi *wordlist-preao-20220621.txt.xz*.