Relatório Final de Atividades - Sistema Atena	
Tipo da Bolsa:	Iniciação Científica
Título do Projeto:	CHATBOT MULTI-AGENTE PARA A DEMOCRACIA DIGITAL
Tópico Abordado:	Aprendizado de Máquina, Processamento de Linguagem Natural e Chatbots
Aluno:	Enzo Bustos da Silva
Unidade do Aluno:	Escola Politécnica
Departamento do Aluno:	PCS
Ano de Ingresso:	2019
Orientador:	Anna Helena Reali Costa
Unidade do Orientador:	Escola Politécnica
Departamento do Orientador:	PCS
Período das Atividades Desenvolvidas:	01/10/2021 a 30/09/2022

1. Principais objetivos iniciais do Plano de Pesquisa:

Dentre os objetivos inicialmente pautados no Plano de Pesquisa estavam os estudos da bibliografia-base em Inteligência Artificial (AI) e Aprendizado de Máquina (AM), além do aprendizado de técnicas específicas para o Processamento de Linguagem Natural (PLN) que envolvem o pré-processamento de textos, extração de informações relevantes e também o uso das ferramentas que são o estado-da-arte atual, como os Transformers e o Hugging Face, usados em diversas aplicações, tais como Sumarização Automática e Modelagem de Tópicos.

2. Descrição das atividades realizadas (mensalmente):

Outubro (2021) - mês 1:

Realização das primeiras etapas da revisão da literatura, focando nos métodos clássicos de AI e AM e aprofundamento na teoria de PLN. Nesta última, citamos os métodos para a vetorização de palavras (Embeddings), de simplificação do léxico (Lematização), tokenizers e stemming. Buscou-se por trabalhos, códigos e bibliotecas que já implementam essas técnicas, a exemplo de bibliotecas temos o SpaCy, Stanza, NLTK, ftfy e gensim. Também foi iniciado um estudo em expressões regulares (RegEx), que é útil para processar textos. Foi utilizado um dataset conhecido de avaliação de filmes, o iMDB dataset para testar em pequena escala tanto alguns desses conceitos como também as bibliotecas que foram citadas. Além disso, foram também pesquisados métodos de extração de textos de PDFs para criar um corpus de dados, testando ferramentas como pdfminer.six, Py2PDF e pdfplumber.

Novembro (2021) - mês 2:

Foi feita a continuação da revisão bibliográfica, atentando principalmente para as publicações mais relevantes e métodos apontados como o estado-da-arte em PLN, como os Transformers, além da revisão da arquitetura de chatbots analisando Intenção, Entidade e Contexto desses agentes. O foco nesta segunda etapa de revisão bibliográfica foi voltado para as pesquisas que envolvem chatbots, estudando publicações sobre os principais conceitos que permeiam a área de chatbots e PLN em geral, envolvendo intenção, entidade e contexto. Para desenvolver habilidades em PLN, o aluno participou do desenvolvimento de um banco de dados em português, para posterior manipulação.

Dezembro (2021) - mês 3:

Foi realizada a criação da base de dados. Os textos utilizados foram extraídos do Diário da Assembleia da República Portuguesa (DAR), utilizando um programa de web scraping baseado em Selenium, de modo a baixar as atas em extensão .txt automaticamente. Após isso foi criada a primeira rotina para mostrar um Dashboard, que consistia de menções diretas ou indiretas à palavra “corrupção”

no discurso dos deputados na DAR. Também foram desenvolvidos os primeiros métodos de segmentação de atas da DAR em blocos de discursos do mesmo assunto, que foi batizado de DEBACER, utilizando Random Forest. No que diz respeito à construção do chatbot, essa mês foi dedicado a revisar quais métodos e técnicas de clustering seriam utilizados para aproximar a construção do chatbot com o projeto corrente do Augmented Democracy. A ideia inicial era criar um FAQ das atas do DAR para que fosse possível que um usuário leigo no assunto pudesse facilmente consultar e tirar dúvidas sobre as atas.

Janeiro (2022) - mês 4:

Foram aplicadas rotinas de pré-processamento ao córpus, como remoção de erros de digitação e stopwords, aplicação de lematização e tokenização, e finalmente, vetorização dos textos utilizando o word2vec pré-treinado. Além da otimização dos algoritmos feitos anteriormente, com a criação também de um notebook interativo (que pode ser visto aqui).

Fevereiro (2022) - mês 5:

O foco principal foi dado à finalização do artigo submetido ao ENIAC 2021 (ver publicações no item 4 deste relatório), em especial no desenvolvimento de um teste de ablação comparando o BERTimbau com outras arquiteturas de AM não-neurais, no contexto do DEBACER. Além disso, foi realizada a apresentação da IC no 29° SIICUSP (link da apresentação) (site do SIICUSP).

O artigo publicado no ENIAC ofereceu um teste de ablação de múltiplos classificadores e também de múltiplas condições dos dados para a tarefa de classificar uma interrupção ou continuidade das falas do Presidente do Diário da Assembleia da República Portuguesa, com a finalidade de separar blocos de falas a respeito de um mesmo assunto/tema. Como dados, usamos as falas do Presidente em diversas configurações, usando diferentes técnicas de extração para features de textos como o Bag of Words, TF-IDF, Bag of N-Grams, Truncated SVD, word2vec, doc2vec, Random Over Sampling e SMOTE, bem como as técnicas de normalização de textos como tokenização, remoção das stopwords e também lematização. No artigo defendemos que apesar do BERTimbau (versão para português do Bidirectional Encoder Representations from Transformers [BERT]) ser uma poderosa técnica no âmbito de NLP, podemos adquirir classificadores estatisticamente equivalentes com modelos muito mais simples, com menor tempo de treinamento e gasto computacional. Para isso testamos KNN, Linear Regression, Random Forest, SGD, SVM e XGBoost. Com uma correta busca de hiperparâmetros, conseguimos, para todas as configurações de dados, modelos que são estatisticamente equivalentes ao BERTimbau utilizando uma análise par a par de Wilcoxon-Holm.

Março (2022) - mês 6:

Neste último mês do semestre de IC, o esforço foi para a submissão de mais dois artigos (ver publicações no item 4 deste relatório):

A. Para o seminário: “INTELIGÊNCIA ARTIFICIAL: DEMOCRACIA E IMPACTOS SOCIAIS” (artigo submetido à Revista do IEA da USP);

B. Para a Conferência Internacional de Processamento Computacional da Língua Portuguesa (artigo aceito no evento PROPOR 2022).

Abril (2022) - mês 7:

Foram feitos estudos e testes com Sumarização Automática, testando diversas arquiteturas como BertExt, BART e PEGASUS. Nos estudos realizados em sumarização automática, investigou-se tanto a sumarização extrativa quanto a abstrativa, em termos gerais e em termos específicos para o projeto com sumarização de diálogos. Dentre as arquiteturas disponíveis para a criação da ferramenta desenvolvida nesta IC, foram escolhidos 3 módulos (ou agentes) para o processamento dos dados do DAR: Modelagem de Tópicos, Sumarização Automática e Análise de Sentimentos. Associados a esses módulos há uma interface e módulos interativos para requisições específicas do usuário.

Maio (2022) - mês 8:

Durante esse período, o bolsista trabalhou para a criação de um algoritmo para criação do algoritmo para sumarizar, sem perda de sentido, as falas de deputados descritas nas Atas do DAR. Para isso foram utilizados alguns modelos para testes. O algoritmo final conta com duas etapas (custosas) de tradução automática e segue o fluxo abaixo:

Fala (pt) → Tradução (pt-en) → Sumarização (en) → Tradução (en-pt) → Resumo

Esta primeira etapa de tradução é bem custosa computacionalmente. Nas etapas de tradução foram testados os modelos PTT5 e M2M100. O primeiro não foi muito bom e precisaria de fine-tuning. O segundo apresentou resultados razoáveis e foi o adotado neste projeto. A etapa de Sumarização foi testada usando os modelos: BertExt, BART, Pegasus e Pegasus-XSUM. O que apresentou melhores resultados foi o modelo Pegasus. A avaliação feita foi subjetiva, utilizando apenas a avaliação humana dos resultados. Mesmo assim, os resultados foram expressivamente superiores aos outros.

Junho (2022) - mês 9:

Foi feita a montagem da funcionalidade de Modelagem de Tópicos, que visa determinar os temas mais relevantes em um documento. A proposta inicial usa o BERTopic para uma primeira segmentação de palavras relacionadas a um mesmo tema e depois atribui o tema final utilizando o modelo roBERTa para Zero-Shot Classification. Os temas finais são definidos pelos projetistas e incluem “Saúde”, “Educação”, “Ciências e Tecnologia”, entre outros. Para a distribuição de um aplicativo web que integre as três funcionalidades desenvolvidas (sumarização,

modelagem de tópicos e análise de sentimentos) o aluno iniciou estudos na ferramenta Django que é integrada com o Python. A ideia foi ter os dados já compilados em um banco de dados para servir principalmente como forma de visualização e interação com o usuário, que poderá aplicar os filtros e restrições desejadas para as visualizações.

Durante este período foi escrito o artigo “ZeroBERTo - Leveraging Zero-Shot Text Classification by Topic Modeling” para a Conferência Internacional de Processamento Computacional da Língua Portuguesa — PROPOR 2022 (ver publicações no item 4 deste relatório).

Julho (2022) - mês 10:

Estudos foram realizados para efetuar o deploy da ferramenta. Inicialmente pensou-se em uma interface web que usa o Django em conjunto com o Python. Foram também estudadas outras abordagens para a criação de chatbots de forma mais automatizada, como o sistema do DiagFlow e Tidio.

Agosto (2022) - mês 11:

Neste período foi definida a arquitetura do chatbot e suas funcionalidades: o chatbot é um robô que responde questões sobre as atas do Diário da Assembleia da República. O robô foi criado através da plataforma DiagFlow e poderá facilmente ser anexado em conjunto com a plataforma web.

Setembro (2022) - mês 12:

Neste último mês foi finalizado o chatbot. A partir da base de dados foi utilizado um gerador de pares de perguntas e respostas (Question Generation) que utiliza o modelo T5. Porém, como o modelo T5 só funciona para a língua inglesa, a base de dados teve que ser traduzida. A partir destes pares foi construída uma base de dados de pares Pergunta e Resposta (Dataset Original). Esta base foi limpa, com a remoção de alguns destes pares por conta principalmente de alucinação do modelo. A alucinação ocorreu por não ter sido usada uma passagem de texto independente e sim um conjunto de falas que têm um contexto. Dessa forma, foram removidas perguntas do tipo “Quem”, que em sua maioria constavam de alucinação do modelo, por conta da perda de contexto e remoção de referências sobre quem estava com a palavra (e a quem estava respondendo), criando-se a nova base de dados (Dataset Limpo). Esta base foi encaminhada ao Knowledge Bases do DialogFlow para que o chatbot pudesse ser montado.

A Figura 1 mostra alguns exemplos da atuação do chatbot. Observe que na primeira (à esquerda), temos um caso que é bem respondido pelo chatbot. Porém, no exemplo da direita percebemos uma alucinação, em que o par de pergunta e resposta foi criado de uma fala que não faz sentido sem o seu contexto.

O Chatbot pode ser testado através deste link.

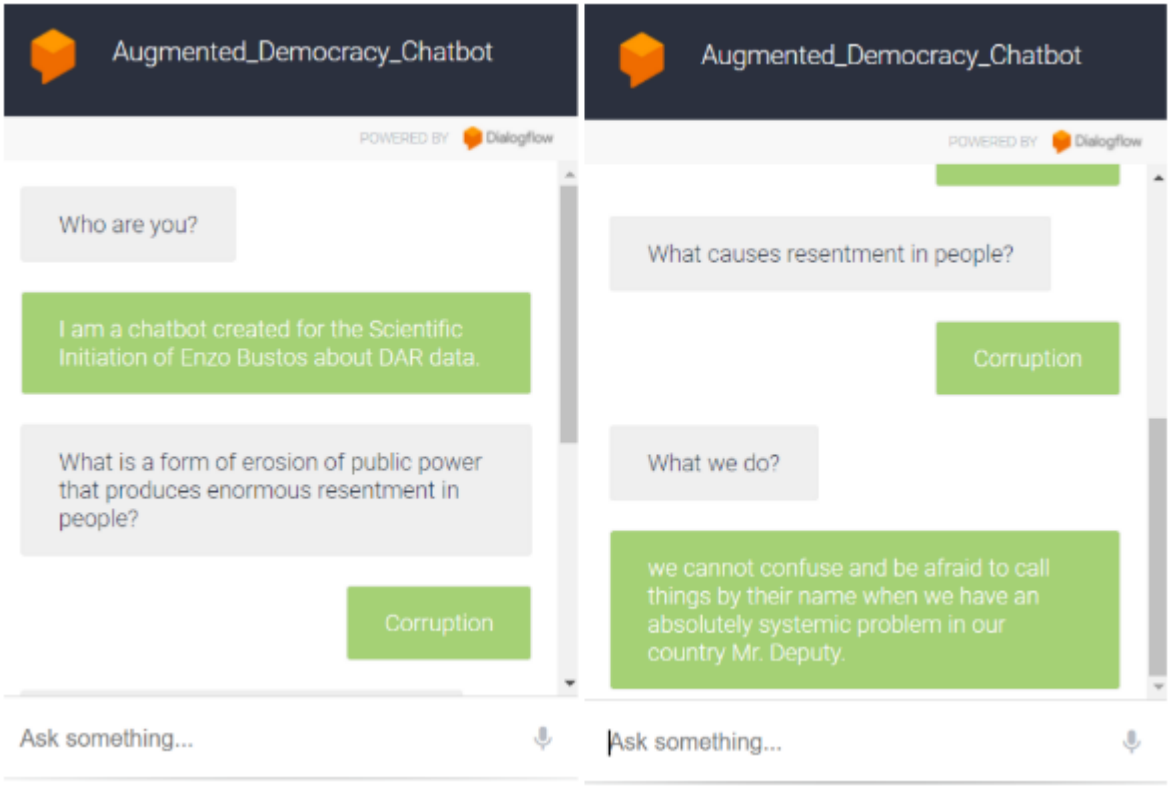

Figura 1 - Exemplos de interação com o chatbot.

3. Principais resultados alcançados:

A construção do chatbot foi finalizada, porém pelo fato do modelo usado não ter sido treinado para textos sequenciais (como é o caso dos textos das atas do DAR, usadas neste trabalho), existem muitos pares que não fazem sentido algum. Outro problema é que a interação com o chatbot atual deve ser feita apenas em inglês por conta do Knowledge Bases, que é uma feature beta do DialogFlow.

Entretanto, do ponto de vista da Iniciação Científica, o aluno aprendeu bastante, fez vários desenvolvimentos computacionais e participou de artigos que foram ou estão sendo publicados. Os principais resultados foram:

- Conclusão de uma boa revisão bibliográfica que serviu para aprofundar os conhecimentos na área de AM e PLN;
- Uso e familiarização com modelos pautados como o estado-da-arte atual;
- Uso da arquitetura de transformers para extrair features mais relevantes dos dados;
- Divulgação Científica;
- Criação do chatbot FAQ sobre as atas do DAR.

4. Artigos publicados:

FERRAZ, Thomas Palmeira et al. DEBACER: a method for slicing moderated debates. *In*: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 18. , 2021, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 667-678. ISSN 2763-9061. DOI: https://doi.org/10.5753/eniac.2021.18293.

Alcoforado, A., Ferraz, T. P., Gerber, R., Bustos, E., Oliveira, A. S., Veloso, B. M., ... & Costa, A. H. R. (2022, March). ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling. In *International Conference on Computational Processing of the Portuguese Language* (pp. 125-136). Springer, Cham.

Publicação de artigo aceito no 1° Seminário Internacional de Humanidades - Artificial Intelligence: Democracy And Social Impacts (C4AI 2021). Revista do IEA da USP: artigo no prelo.

Apresentação da IC no 29° SIICUSP:
Link da apresentação: SIICUSP - Sala 10
Acesso à publicação: "Democracia Aumentada: Um Sistema para Democracia Digital", edição de 2021 (site do SIICUSP).

Democracia Aumentada: Um Sistema para Democracia Digital*

Enzo Bustos da Silva
Thomas Palmeira Ferraz, André Seidel Oliveira
Anna Helena Reali Costa
Escola Politécnica da Universidade de São Paulo
enzobustos@usp.br

Objetivos

Com a crescente disseminação da Internet, a participação democrática tem sido aprimorada tanto no processamento de informações, como na comunicação e transações. Por exemplo, o cidadão pode obter informações sobre a política local e avaliar a atuação de seus eleitos, além de nortear seus votos e sua atuação política. Este trabalho propõe a Democracia Aumentada, um sistema aplicado à democracia digital [1] que emprega técnicas de aprendizado de máquina (AM) e processamento de linguagem natural (PLN) visando aumentar a transparência do processo democrático. Isso foi feito através da interpretação automática dos documentos gerados em órgãos públicos, com a finalidade de traduzir esses textos, extensos e de difícil interpretação, em algo que seja mais facilmente compreendido por qualquer cidadão.

Mais especificamente, técnicas de AM e PLN foram investigadas nas atas do Diário da Assembleia da República Portuguesa[1] almejando três contribuições:

1. Processar e estruturar os dados produzidos pelo corpo legislativo;
2. Extrair e processar informações relevantes dentro do discurso político – em especial, com sumarização automática, modelagem de tópicos e análise de sentimentos;
3. Providenciar formas de interpretação, visualização e interação que facilite o entendimento do cidadão comum.

Métodos e Procedimentos

Para a criação deste sistema, foi necessário primeiramente usar métodos de *web crawling* e *web scraping* para coletar os dados das atas diretamente do *website* do Parlamento Português. Estes dados crus passam então por uma pré-segmentação de modo a dividir o texto inteiro do conjunto de Atas em uma base de dados estruturada, como mostrado na Figura 1, facilitando assim o processamento e a extração de informações a partir destes dados.

	Transcript	Date	Person	Party	Text	Initiatives
498	DAR-002	18 DE SETEMBRO DE 2020	André Ventura	CH	Sr. Presidente, Srs. Deputados: Temos de começ...	projeto de resolução n.º 471, projeto de lei n...
499	DAR-002	18 DE SETEMBRO DE 2020	Presidente	Fernando Negrão	Sr. Deputado, chamo a sua atenção para o tempo...	projeto de resolução n.º 471, projeto de lei n...
500	DAR-002	18 DE SETEMBRO DE 2020	André Ventura	CH	Vou terminar, Sr. Presidente. Portanto, não há...	projeto de resolução n.º 471, projeto de lei n...
501	DAR-002	18 DE SETEMBRO DE 2020	Duarte Alves	PCP	Quando pagamos!...	SÉRIE — NÚMERO 2	projeto de resolução n.º 471, projeto de lei n...
502	DAR-002	18 DE SETEMBRO DE 2020	André Ventura	CH	Se pesquisarmos na internet por PCP e nacional...	projeto de resolução n.º 471, projeto de lei n...
503	DAR-002	18 DE SETEMBRO DE 2020	Presidente	Fernando Negrão	Tem de terminar, Sr. Deputado.	projeto de resolução n.º 471, projeto de lei n...

Figura 1: Dados gerados a partir das atas do DAR.

Nesses dados, vale ressaltar as colunas “Text” e “Initiatives”. A primeira corresponde a cada uma das falas que ocorreram na sessão plenária, divididas por locutor; a segunda trata de qual item de pauta corresponde aquela fala (por exemplo, sobre um Projeto de Resolução).

A partir dessa estruturação, as interações do usuário com o sistema seguem o esquema da Figura 2. O usuário insere informações que lhe são pertinentes, como o período de interesse, um partido ou político específico ou ainda um determinado assunto de interesse.

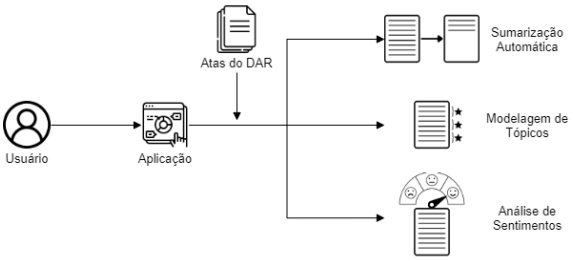

Figura 2: Esquema do sistema de Democracia Aumentada proposta, com funcionalidades de sumarização automática, definição do tópico abordado e análise dos sentimentos das falas.

Em conjunto com as informações fornecidas pelo usuário, algoritmos de AM sofisticados são aplicados, como BERT [2] e suas variantes do estado-da-arte, visando realizar três funcionalidades especializadas:

1. Sumarização Automática [3]: visa produzir um resumo a partir de um conjunto de documentos de entrada;
2. Modelagem de Tópicos [4]: método para identificar conjuntos de palavras para determinar o tópico abordado em um texto;
3. Análise de Sentimentos [5]: tarefa que almeja determinar a polaridade do sentimento expresso em um texto.

Para esse trabalho, utilizamos (1) para reduzir o tamanho das discussões de um determinado item de pauta; (2) como uma ferramenta de busca, orientando o usuário para discussões que são de seu interesse e (3) para certificar que os argumentos apresentados por um partido são condizentes com seu voto. O trabalho ainda está em andamento, com previsão de término em fevereiro de 2022.

Resultados

Como resultado desse projeto de iniciação científica temos um sistema unificado que engloba diversas funcionalidades de AM e PLN aplicadas no contexto das Atas do Diário da Assembleia da República Portuguesa.

Um dos resultados parciais pode ser visto na Figura 3, em que o algoritmo da modelagem de tópicos encontra as principais palavras que estão relacionadas a um mesmo assunto.

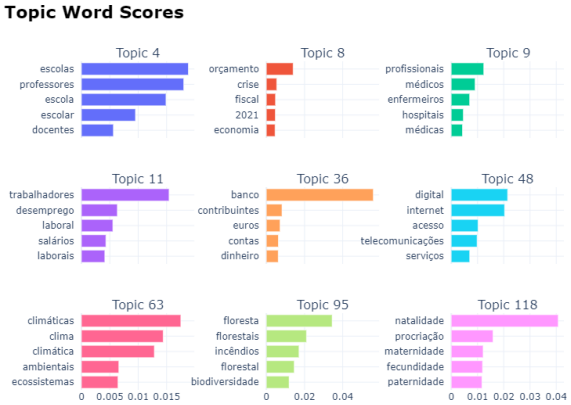

Figura 3: Exemplo dos principais tópicos identificados pelo BERTopic nas atas.

Outro resultado, agora da sumarização automática, pode ser visto na Figura 4, no qual um texto de intervenção[2] de 7903 caracteres é reduzido para apenas um parágrafo sem grandes perdas no sentido da fala completa.

> O deputado Antnio Filipe, com correção, disse que este debate nem se formulou como um debate sobre corrupção neste contexto, ou sobre os líderes que são acusados de corrupção, ou sobre os líderes que são efetivamente condenados por corrupção, porque transformá-lo, como foi transformado, em um debate sobre suspeitas amplas é tudo que não podemos fazer na democracia.

Figura 4: Exemplo de um sumário automático gerado pelo PEGASUS usando técnicas abstrativas.

Conclusões

O sistema de Democracia Aumentada proposto neste trabalho contribui especialmente para o primeiro eixo da democracia digital, a informação, demonstrando que a integração de técnicas de inteligência artificial no âmbito político pode melhorar a qualidade da informação, tornando-a mais concisa e direta, para ser usufruída pela sociedade.

Além de uma base para outros projetos que podem desenvolver os demais eixos da democracia digital, de discussão e participação, planejamos construir uma interface amigável que englobe todas essas funcionalidades para o usuário final.

Referências Bibliográficas

[1] Breindl, Yana et al. "Can Web 2.0 applications save e-democracy? A study of how new internet applications may enhance citizen participation in the political process online". International Journal of Electronic Democracy 1. 1(2008): 14–31.

[2] Devlin, Jacob et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805. (2018).

[3] Huang, Dandan et al. "What Have We Achieved on Text Summarization?". arXiv preprint arXiv:2010.04529. (2020).

[4] Maarten Grootendorst. (2020). BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.

[5] Abercrombie, Gavin, and Riza Theresa Batista-Navarro. "'Aye' or 'no'? Speech-level sentiment analysis of Hansard UK parliamentary debate transcripts." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[1] Disponível em: https://www.parlamento.pt/DAR
[2] Intervenção na íntegra: https://debates.parlamento.pt/catalogo/r3/dar/01/14/02

*Agradecimentos ao Programa de Bolsas Itaú em conjunto com o C²D pelo financiamento parcial deste projeto.