UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PROCESSAMENTO DE LINGUAGEM NATURAL

Prof.: Arlindo Galvão

Dyonnatan Ferreira Maia
João Gabriel Junqueira

Fevereiro de 2023

Relatório

Detecção de Posicionamento na opinião do cidadão

Resumo

O relatório apresenta o estudo de caso como trabalho final da disciplina de Processamento de Linguagem Natural ofertada no Instituto de Informática da UFG. O trabalho proposto consiste na entrega de dois estudos pilotos, o primeiro relacionado à tarefa de detecção de posicionamento em comentários dos cidadãos sobre os Projetos de Lei da Câmara dos Deputados, e o segundo estudo tem como objetivo unir diversas tarefas de NLP (similaridade semântica, agrupamento de texto, classificação, modelagem de tópicos e ranqueamento) à fim de entregar um produto que permita o usuário pesquisar comentários de acordo com o posicionamento dentro de uma base de dados. Como resultado foram obtidos os modelo de classificação de posicionamento mBERT (F1=0,981 e Loss-validação=0,951) e BERTimbau (F1=0,946 e Loss-validação=0,891), para a segunda etapa do projeto foi obtido como resultado um Jupyter Notebook com todas as tarefas na pipeline de execução e entregando como saída os top comentários posicionados de acordo com o tema e posicionamento escolhidos pelo usuário.

Introdução

A detecção de posicionamento trata da identificação e classificação do posicionamento expresso por um indivíduo sobre algum tema ou tópico, isto é, se o texto apresenta pontos de vista ou argumentos concordando ou não com um determinado tema. A detecção de posicionamento se diferencia da análise de sentimentos em dois aspectos principais, primeiramente possuem objetivos distintos como pode ser observada na frase "infelizmente as pessoas não se vacinaram", a frase possui um sentimento negativo contudo um posicionamento positivo com relação à vacina. O segundo aspecto refere-se à estrutura da tarefa, o sentimento detectado pode ser de forma geral olhando a frase como um todo, porém o posicionamento exige como informação de entrada a tupla frase e tema pois a detecção é sempre relativa à algo.

Para o desenvolvimento da sociedade é importante compreender os posicionamentos da população, porém em um alto volume de dados gerados se torna desafiador ouvir o cidadão. No escopo político, quando um novo Projeto de Lei (PL) recebe maior atenção do público, isso faz com que as pessoas discutam o conteúdo e expressem seus pontos de vista mais ativamente na internet. Com a detecção de posicionamento é possível automatizar e reduzir o viés do trabalho manual de leitura e anotação desses dados, permitindo uma análise mais acurada pelos interessados no desenvolvimento do tema, como os próprios responsáveis pela aprovação dos projetos.

Com isso, o presente projeto propõe um estudo sobre a detecção de posicionamento aplicado a textos curtos referentes a comentários públicos sobre Projetos de Lei da Câmara dos Deputados do Brasil.

Formulação do problema

No site da Câmara dos Deputados do Brasil há uma seção de enquetes [1] sobre os Projetos de Lei a fim de ouvir a opinião do cidadão, servindo de ferramenta para compreensão da aceitação do projeto e identificação de possíveis melhorias para o texto. Porém, os recursos utilizados na enquete são insuficientes para identificar os comentários favoráveis e contrários ao projeto de lei, sendo necessária uma leitura ostensiva pela equipe da Câmara para classificação dos comentários. Com isso, é proposto a classificação automática dos comentários de acordo com o posicionamento no tema. Para tal tarefa, há alguns desafios a serem superados, a falta de um corpus suficientemente grande para a língua portuguesa ou um corpus específico para o contexto especificado. Considerando que um deputado pode apresentar um projeto sobre qualquer tema e a relevância em analisar esses posicionamentos duram até o dia da votação sendo substituídos pelo interesse nos próximos temas a serem votados, devido a essa característica, a demanda dos temas se tornam imprevisíveis e possuem um tempo de vida relativamente curto, com isso, é interessante a análise da capacidade do modelo em suprir essa demanda variável de tópicos, dado que boa parte dos estudos fixam alguns poucos temas arbitrariamente escolhidos para os experimentos.

Com objetivo de no fim deste projeto obter um protótipo da aplicação também há a necessidade de explorar outras técnicas de Processamento de Linguagem Natural para que

[1] https://www.camara.leg.br/enquetes

retorne comentários relevantes da base de dados de acordo com a consulta realizada pelo usuário. Para isso, pode-se realizar uma busca por comentários candidatos, filtrar os comentários que possuam os posicionamentos de interesse, e retornar o resultado ranqueado pela relevância de cada comentário e como informação complementar retornar a identificação dos principais tópicos citados nestes comentários.

Metodologia

O estudo foi dividido em três partes: coleta e tratamento dos dados para construção do córpus, realização dos experimentos e construção do protótipo de aplicação. Na primeira parte, foram coletados os comentários públicos dos usuários da seção de enquete de projetos no site da Câmara dos Deputados (Figura 1) [4], também foram implementados os modelos propostos no estudo de Pavan et al. [1, 2] e comparado com o modelo proposto. Na segunda parte foram realizados experimentos relativos a cada tarefa proposta, sendo elas: detecção de posicionamento, busca de informação com similaridade semântica, ranqueamento de comentários com BM25, e modelagem de tópicos. Na terceira parte, foi construído com Jupyter Notebook, utilizando os melhores modelos obtidos para realizar as inferências, um *pipeline* que simula a interação do usuário e retorna as informações relativas à busca.

Figura 1: Comentários da página de enquetes da Câmara dos Deputados.

Os dados obtidos da coleta foram anotados por três grupos composto por três pessoas, cuja concordância foi de (Kappa k_1 = 0,47, k_2 = 0.31 e k_3 = 0.62, respectivamente). De acordo com a Tabela 1 foram aproximadamente 80% de dados para treino e 20% para teste. Sendo os tópicos de treino: Desarmamento, Servidores Públicos, Contratação, Código Penal, Estatuto do Desarmamento, Reforma Administrativa, Reforma Tributária, CLT, Reforma Trabalhista, Ajuda de custo, Reforma Previdenciária, Partidos Políticos, Seguro-Desemprego, Porte de Armas, Estatuto

da OAB, Salário Mínimo, LDB, Lei Maria da Penha e Código de Defesa do Consumidor. E os tópicos de teste: Ajuda de Custo (AC), CLT, LOAS e Servidores Públicos (SPs) [3].

Conjunto	Favorável	Contrário	Nenhum	Total
Treino	530	825	147	1502
Teste	228	137	69	434
Total	758	962	216	1936

Tabela 1: Dados de treinamento para o modelo de Detecção de Posicionamento

Para a terceira etapa, o desenvolvimento do protótipo funcional foi definido de acordo com os seguintes descritos na Figura 2.

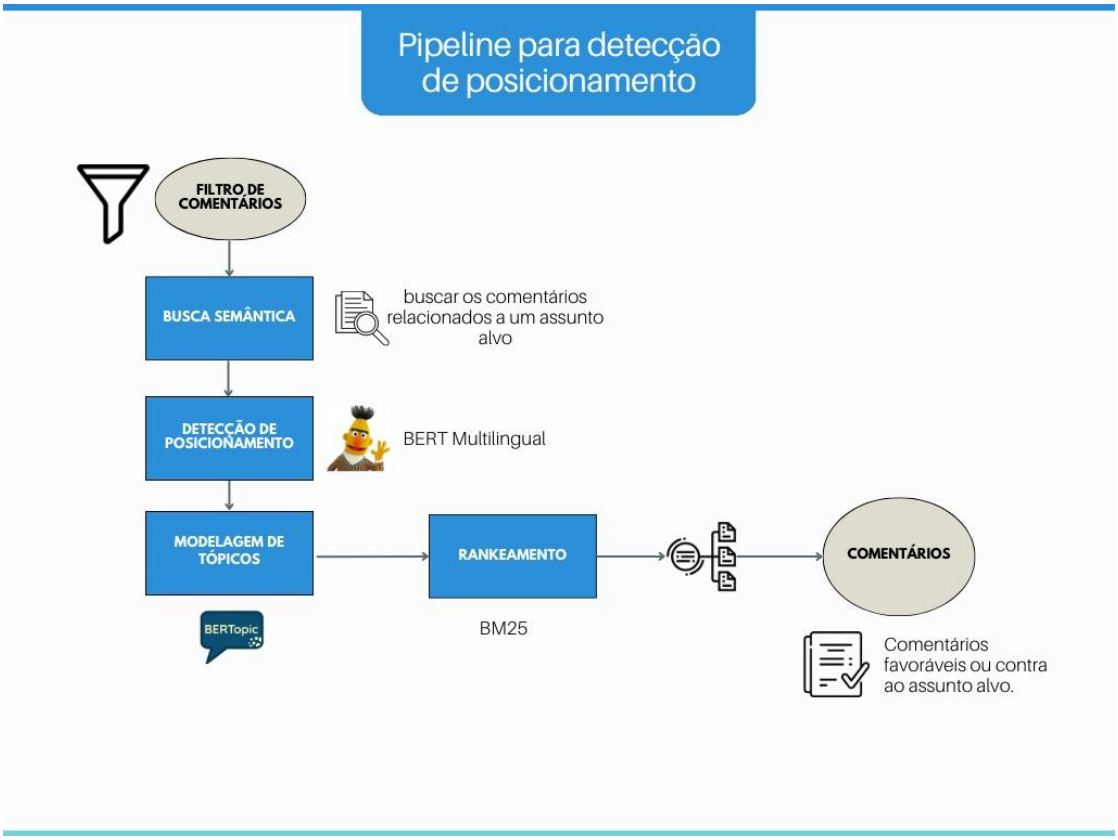


Figura 2: Pipeline do protótipo funcional

Resultados e discussões

- Detecção de Posicionamento

O modelo proposto é composto por um modelo BERT pré-treinado com duas camadas FeedForward de (768, 2) neurônios, a entrada é composta pelo par de sentença (<CLS>tema<SEP>comentário<SEP>). Foram treinados e avaliados dois modelos pré-treinados para a classificação do posicionamento, o BERTimbau e BERT multilíngue (mBERT). Também foram adicionados no treinamento os algoritmos propostos por Pavan et. al [1] para comparação.

Tópico	NB	RL	MLP	SVM	RF	BERTimbau	mBERT	Qtd.
AC	0.779	0.806	0.777	0.762	0.742	**0.904**	0.887	44
CLT	0.817	0.656	0.522	0.560	0.573	0.901	**1.000**	72
LOAS	0.359	0.521	0.557	0.557	0.585	0.875	**0.889**	18
SPs	0.615	0.503	0.349	0.430	0.281	0.973	**1.000**	231

Tabela 2: F1 ponderado para classificação de posicionamento nos dados da Câmara.

Como pode ser observado na Tabela 2 o mBERT atingiu os melhores resultados com a média ponderada dos F1 igual a 0,981 enquanto que o BERTimbau atingiu F1 = 0,946. Contudo, comparando a curva de loss de ambos os modelos (Figuras 3 e 4) o BERTimbau apresentou menor viés aos dados de treinamento, a curva de validação se refere à 30% dos dados de treino.

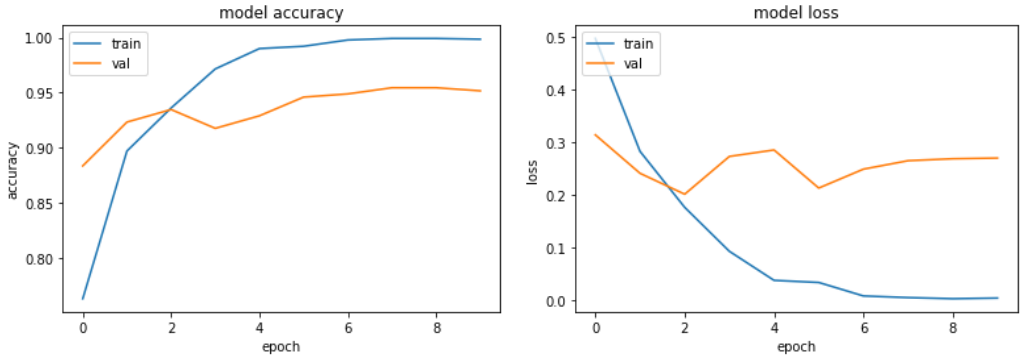


Figura 3: Curvas de acurácia e loss do modelo mBERT

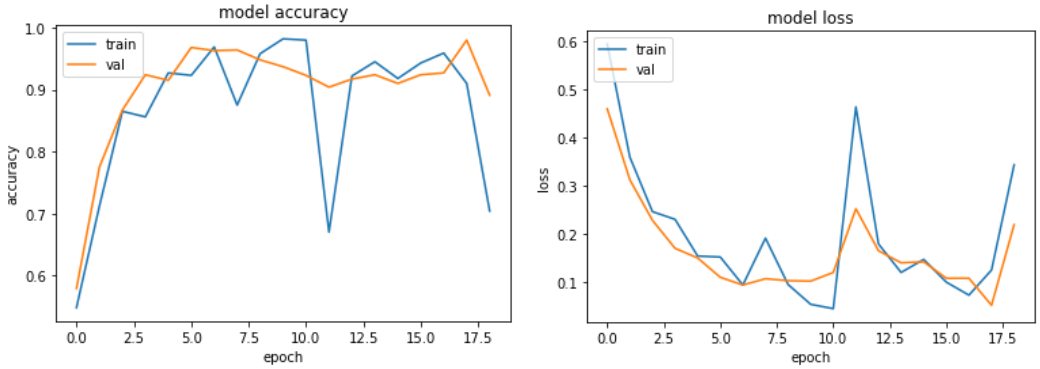


Figura 4: Curvas de acurácia e loss do modelo BERTimbau

- Protótipo

Para o desenvolvimento da aplicação, algumas técnicas de NLP foram adicionadas à pipeline de demonstração.

Na primeira etapa, para realizar uma busca otimizada aos dados que serão classificados, foi aplicado uma indexação dos comentário através de embeddings gerados pelo modelo de sentença pré-treinado paraphrase-multilingual-mpnet-base-v2 para realizar uma filtragem por similaridade semântica com o tema alvo escolhido pelo usuário. A seleção do modelo foi feita através de uma análise humana, onde os demais modelos avaliados foram: msmarco-Mini LM-L-12-v3 e distilbert-dot-tas_b-b256-msmarco. A indexação com armazenamento feito na memória foi realizada com a biblioteca Faiss. Como estudo piloto, a fim de priorizar comentários onde facilmente identifica-se os temas citados, foi aplicada uma filtragem com cálculo de similaridade através da distância Levenshtein implementada com RapidFuzz, assim aumenta a probabilidade de ler comentários que contenham explicitamente o tema de interesse. O próximo passo foi aplicar o modelo BERTimbau com o finetune deste domínio para tarefa de detecção de posicionamento para realizar as inferências de posicionamento nos comentários filtrados. Para visualização dos resultados foi aplicado o algoritmo BM25 para ranquear os comentários de acordo com frequencia dos tokens e dos documentos/comentários normalizados pelo comprimento de cada documento, assim serão lidos primeiro os comentários que podem trazer um vocabulário que utilize as palavras relevantes mais comuns dentre os comentários posicionados. Para concluir a pipeline, a última tarefa adicionada tem como objetivo pontuar quais outros tópicos são citados dentre os comentários para que possa auxiliar o entendimento de porque os comentários possuem tais posicionamentos e até mesmo sirva de direcionamento para encontrar outros tópicos relevantes para a análise do posicionamento do cidadão. Para última tarefa foi realizada modelagem de tópicos para encontrar novos tópicos dentre os comentários que possuem o posicionamento para o tema escolhido, utilizando a biblioteca BERTopic [5] foram montados os seguintes passos: geração de embeddings com o modelo pré-treinado multilingual SentenceBERT, redução de dimensionalidade com UMAP, clusterização dos vetores com HDBSCAN, tokenização com CountVectorizer, extração dos tópicos com class-based TF-IDF para pegar as palavras “mais importantes” de cada cluster.

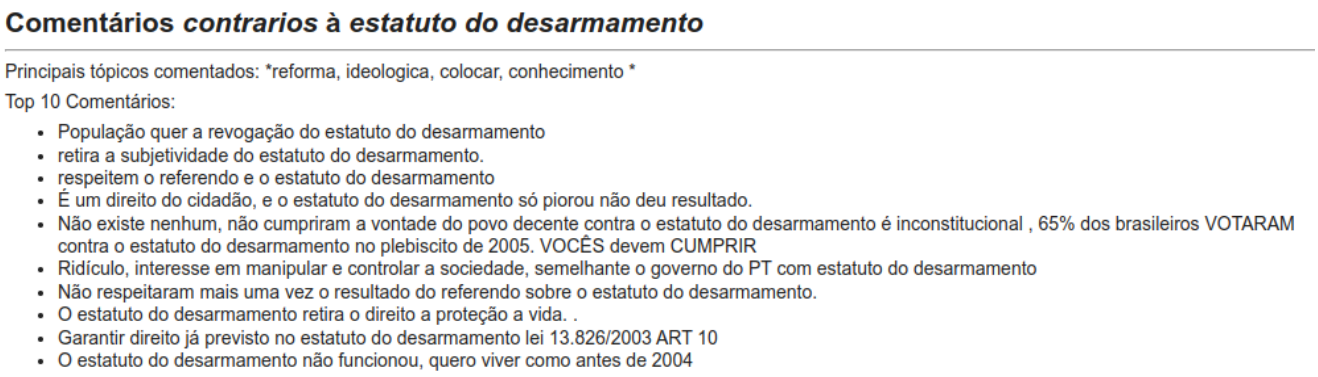
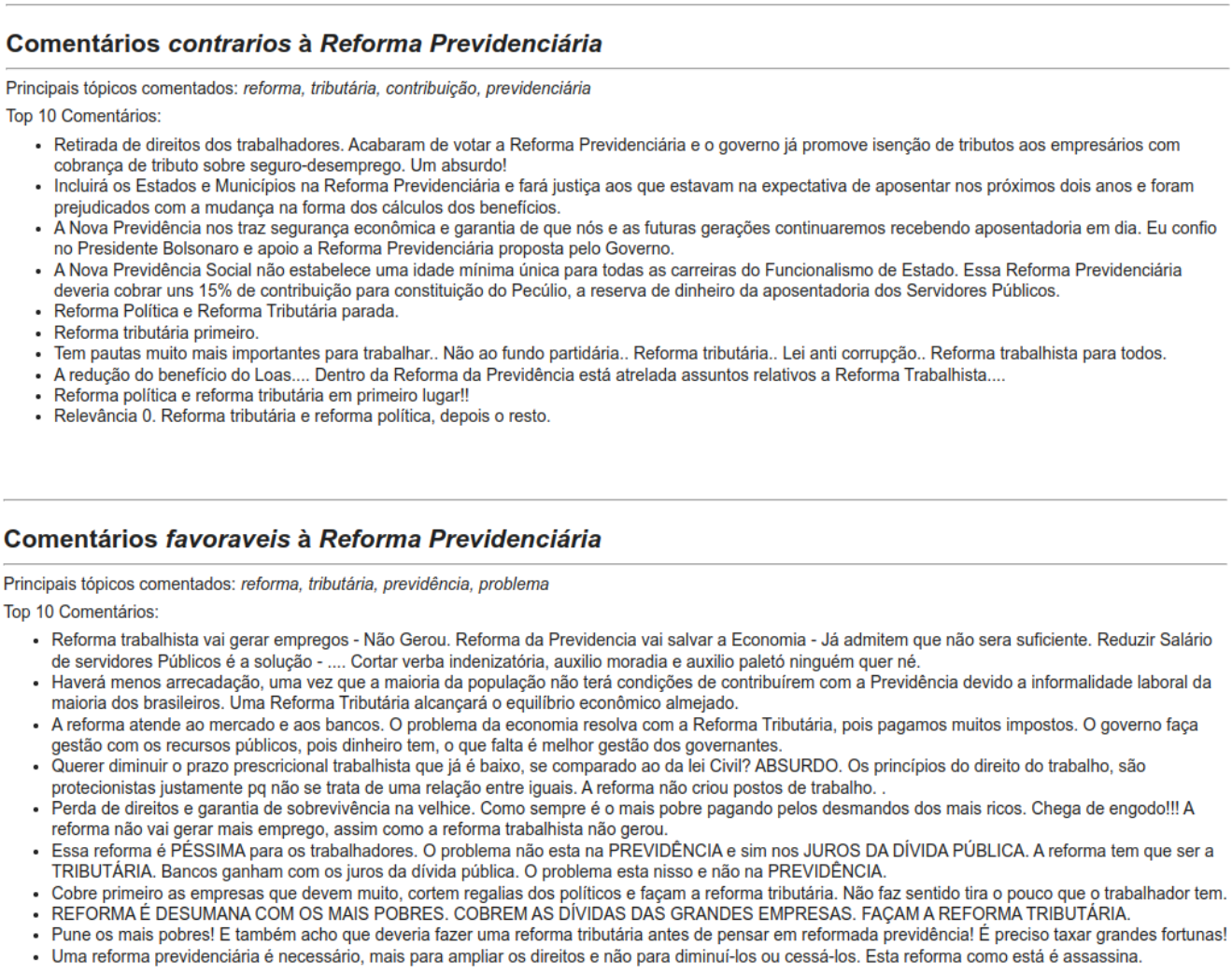


Figura 5: Saída da pipeline do protótipo proposto. Top 10 comentários considerados contrários ao estatuto do desarmamento.

O resultado final pode ser visto nas Figuras 5 e 6, onde o usuário passa o tema e tipo de posicionamento (contrário ou favorável) como entrada, e retorna os comentários de acordo com o posicionamento escolhido, ranqueados de acordo com relevância das palavras e os tópicos citados nestes comentários.



Comentários *contrarios* à *Reforma Previdenciária*

Principais tópicos comentados: *reforma, tributária, contribuição, previdenciária*

Top 10 Comentários:

- Retirada de direitos dos trabalhadores. Acabaram de votar a Reforma Previdenciária e o governo já promove isenção de tributos aos empresários com cobrança de tributo sobre seguro-desemprego. Um absurdo!
- Incluirá os Estados e Municípios na Reforma Previdenciária e fará justiça aos que estavam na expectativa de aposentar nos próximos dois anos e foram prejudicados com a mudança na forma dos cálculos dos benefícios.
- A Nova Previdência nos traz segurança econômica e garantia de que nós e as futuras gerações continuaremos recebendo aposentadoria em dia. Eu confio no Presidente Bolsonaro e apoio a Reforma Previdenciária proposta pelo Governo.
- A Nova Previdência Social não estabelece uma idade mínima única para todas as carreiras do Funcionalismo de Estado. Essa Reforma Previdenciária deveria cobrar uns 15% de contribuição para constituição do Pecúlio, a reserva de dinheiro da aposentadoria dos Servidores Públicos.
- Reforma Política e Reforma Tributária parada.
- Reforma tributária primeiro.
- Tem pautas muito mais importantes para trabalhar.. Não ao fundo partidária.. Reforma tributária.. Lei anti corrupção.. Reforma trabalhista para todos.
- A redução do benefício do Loas.... Dentro da Reforma da Previdência está atrelada assuntos relativos a Reforma Trabalhista....
- Reforma política e reforma tributária em primeiro lugar!!
- Relevância 0. Reforma tributária e reforma política, depois o resto.

Comentários *favoraveis* à *Reforma Previdenciária*

Principais tópicos comentados: *reforma, tributária, previdência, problema*

Top 10 Comentários:

- Reforma trabalhista vai gerar empregos - Não Gerou. Reforma da Previdencia vai salvar a Economia - Já admitem que não sera suficiente. Reduzir Salário de servidores Públicos é a solução - .... Cortar verba indenizatória, auxilio moradia e auxilio paletó ninguém quer né.
- Haverá menos arrecadação, uma vez que a maioria da população não terá condições de contribuírem com a Previdência devido a informalidade laboral da maioria dos brasileiros. Uma Reforma Tributária alcançará o equilíbrio econômico almejado.
- A reforma atende ao mercado e aos bancos. O problema da economia resolva com a Reforma Tributária, pois pagamos muitos impostos. O governo faça gestão com os recursos públicos, pois dinheiro tem, o que falta é melhor gestão dos governantes.
- Querer diminuir o prazo prescricional trabalhista que já é baixo, se comparado ao da lei Civil? ABSURDO. Os princípios do direito do trabalho, são protecionistas justamente pq não se trata de uma relação entre iguais. A reforma não criou postos de trabalho. .
- Perda de direitos e garantia de sobrevivência na velhice. Como sempre é o mais pobre pagando pelos desmandos dos mais ricos. Chega de engodo!!! A reforma não vai gerar mais emprego, assim como a reforma trabalhista não gerou.
- Essa reforma é PÉSSIMA para os trabalhadores. O problema não esta na PREVIDÊNCIA e sim nos JUROS DA DÍVIDA PÚBLICA. A reforma tem que ser a TRIBUTÁRIA. Bancos ganham com os juros da dívida pública. O problema esta nisso e não na PREVIDÊNCIA.
- Cobre primeiro as empresas que devem muito, cortem regalias dos políticos e façam a reforma tributária. Não faz sentido tira o pouco que o trabalhador tem.
- REFORMA É DESUMANA COM OS MAIS POBRES. COBREM AS DÍVIDAS DAS GRANDES EMPRESAS. FAÇAM A REFORMA TRIBUTÁRIA.
- Pune os mais pobres! E também acho que deveria fazer uma reforma tributária antes de pensar em reformada previdência! É preciso taxar grandes fortunas!
- Uma reforma previdenciária é necessário, mais para ampliar os direitos e não para diminuí-los ou cessá-los. Esta reforma como está é assassina.

Figura 6: Saída da pipeline do protótipo proposto. Top 10 comentários considerados contrários ao estatuto do desarmamento.

Conclusão

O relatório tratou da apresentação final dos experimentos realizados sobre o estudo de caso para detecção de posicionamento na opinião do cidadão e elaboração de um projeto piloto para disponibilizar uma interface para o usuário realizar pesquisas de comentário de acordo com o posicionamento. Na primeira etapa foi realizada a coleta e tratamento dos dados, resultando em um dataset dividido em aproximadamente com 20% para teste e 80% dos dados para treinamento ao qual foi subdividido em treino (70%) e validação (30%). Os modelos treinados BERTimbau e mBERT superaram os modelos utilizados de trabalhos anteriores (NB, RL, MLP, SVM e RF) para a classificação de posicionamento. A última etapa foi a construção do pipeline para testar um protótipo funcional onde foram adicionadas similaridade semântica que obteve resultados interessantes porém podem melhorar em trabalhos futuros, o ranqueamento com BM25 que também apresentou resultados visualmente interessantes e que em trabalhos futuros podem também ser adicionados no ranqueamento a nota dos usuários para o comentário, para a modelagem de tópicos, devido a pouca quantidade de comentários em alguns casos mostrou que cabe melhorias nas escolhas destes algoritmos. Porém, por fim, o resultado final se mostrou interessante e que já poderia ser utilizado para auxiliar na leitura e entendimento da opinião do cidadão, visto que os top comentários exibidos estão relacionados e apresentam conteúdos relevantes para melhor compreender os anseios da população e direcionar as políticas públicas.

Referências

1. Pavan, M.C., Dos Santos, V.G., Lan, A.G.J., Martins, J., Santos, W.R., Deutsch, C., Costa, P.B., Hsieh, F.C., Paraboni, I.: Morality classification in natural language text. **IEEE Transactions on Affective Computing** pp. 1–1 (2020). https://doi.org/10.1109/TAFFC.2020.3034050.
2. Pavan, M., dos Santos, W., Paraboni, I.: Twitter moral stance classification using long short-term memory networks. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)** 12319 LNAI, 636–647 (2020). https://doi.org/10.1007/978-3-030-61377-8-45.
3. KÜÇÜK, D.; CAN, F. Stance detection: A survey. **ACM Comput**. Surv. 53(1), feb 2020.
4. Maia, D.F. et al. (2022). UlyssesSD-Br: Stance Detection in Brazilian Political Polls. In: Marreiros, G., Martins, B., Paiva, A., Ribeiro, B., Sardinha, A. (eds) Progress in Artificial Intelligence. EPIA 2022. Lecture Notes in Computer Science(), vol 13566. Springer, Cham. https://doi.org/10.1007/978-3-031-16474-3_8
5. GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. **arXiv preprint** arXiv:2203.05794, 2022.