Análise de Discursos de Deputada(o)s Brasileiros via Modelos de Linguagem

Pedro Dalla Vecchia Chaves
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG
pedrodallav@dcc.ufmg.br

1 INTRODUÇÃO

Praticamente todos os dias deputadas e deputados de diferentes partidos políticos proferem discursos e pronunciamentos na Câmara dos Deputados sobre os mais diversos assuntos. A transcrição bem como um sumário de tais discursos podem ser acessados, lidos e analisados por qualquer pessoa através do Portal da Câmara (link).

Analisar o teor de discursos políticos, em termos de tentar identificar padrões de concordância/relação entre discursos, é uma tarefa complexa do ponto de vista humano, visto a possibilidade da introdução de vieses de interpretação de cada indivíduo.

Uma possível tentativa para "formalizar"esse processo seria através de modelos de linguagem, especialmente os recentes modelos de linguagem neural, que são capazes de aprender representações vetoriais de palavras e sentenças em um espaço semântico latente. Após o treinamento de tais modelos, podemos comparar diferentes discursos através de operações que envolvem calcular a distância entre os vetores das palavras presentes nesses discursos, a fim de obter uma noção de concordância/relação.

O objetivo do trabalho em questão é aplicar diferentes modelos de linguagem neurais sobre os discursos de deputadas e deputados de diferentes partidos a fim de se tentar construir uma noção de relação entre tais discursos, observando se são próximos no que diz respeito ao espectro político esquerda-direita tão mencionado nos dias atuais.

2 RERENCIAL TEÓRICO

A seguir serão apresentados tópicos e conceitos relacionados à implementação do trabalho.

2.1 Modelos de linguagem probabilísticos

Modelos de linguagem estatísticos/probabilísticos [15] são modelos paramétricos baseados em distribuições de probabilidade sobre sequências de palavras. Essas probabilidades podem ser calculadas com base na ocorrência absoluta (contagem) de cada uma das palavras em um corpus ou na co-ocorrência com outras palavras nesse corpus (n-grams).

Entretanto, tais modelos apresentam limitações no que diz respeito à "cobertura"dos dados, podendo gerar combinações/co-ocorrências de palavras de forma inválida e possivelmente bastante numerosa, visto a possibilidade de distribuições conjuntas. Além disso, as representações utilizadas não são capazes de capturar sentidos latentes das palavras nas sentenças.

2.2 Modelos de linguagem neurais

Modelos de linguagem neurais [4] são uma espécie de evolução dos modelos de linguagem probabilísticos que tentam aprender uma representação distribuída das palavras/sentenças de um corpus a partir da utilização de certas arquiteturas de redes neurais, tentando resolver o problema relacionado à *curse of dimensionality* gerado nos modelos probabilísticos.

Avanços mais recentes em modelos de linguagem neurais [13] [12] permitiram a criação de representações vetoriais distribuídas mais robustas, capazes de capturar noções semânticas das palavras. Tais modelos, comumente chamados de Word2Vec apresentam duas variantes principais:

- CBOW: arquitetura de rede neural que tenta predizer a próxima palavra de uma sentença com base no contexto.
- Skip-Gram: arquitetura de rede neural que tenta predizer as palavras em volta de uma palavra alvo.

Ao tentar predizer o contexto ou a próxima palavra, tais redes vão aprendendo as representações vetoriais distribuídas.

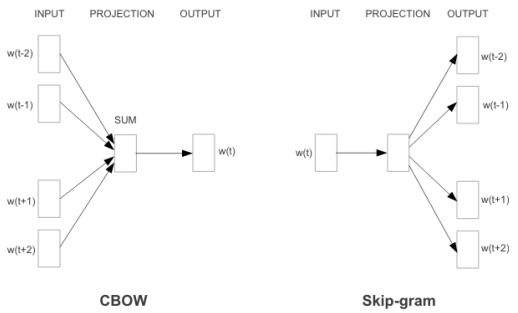

Figura 1: Representação visual do funcionamento das arquiteturas CBOW e Skip-Gram. Retirado de [12]

2.3 Word Mover's Distance (WMD)

Nas áreas de Recuperação de Informação e Sistemas de Recomendação normalmente são utilizadas diferentes métricas para calcular a distância/similaridade entre documentos [8], que nada mais são do que conjunto de sentenças/palavras. Tais métricas resumem-se em cálculos de distância com base em vetores de características (ou *features*), como por exemplo vetores que representam a frequência de ocorrência de cada uma das palavras do documento.

Para tirar proveito das representações vetoriais distribuídas explicadas anteriormente, foi desenvolvida uma métrica capaz de

mensurar a distância entre um conjunto dessas representações. Tal métrica é chamada de Word Mover's Distance (WMD) [10].

A métrica tenta representar a distância acumulativa mínima para "transportar"todos os vetores de palavras de uma sentença para o espaço ocupado pelos vetores de palavras de outra sentença.

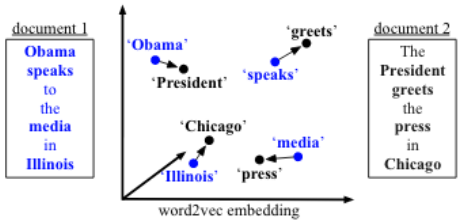

Figura 2: Representação visual do funcionamento do algoritmo que calcula a WMD entre duas sentenças. Retirado de [10]

2.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Normalmente as representações distribuídas aprendidas apresentam dimensionalidade relativamente alta, variando de 50 até 500 dimensões. Visualizar proximidades entre tais vetores se torna inviável quando estamos em dimensões maiores que 3. Para conseguirmos visualizar é necessário reduzir a dimensão dos vetores via algum tipo de algoritmo de redução de dimensionalidade.

Existem inúmeros tipos de algoritmos para realizar a tarefa de redução [16]. Um método relativamente recente e comumente utilizado para visualizar representações de alta dimensão é o t-SNE [11]. O funcionamento da redução de dimensionalidade via t-SNE consiste em tentar manter as relações espaciais presentes em espaços n dimensionais ao se reduzir para espaços k dimensionais, onde $k < n$.

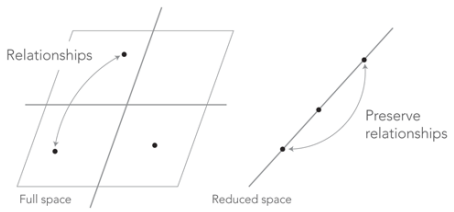

Figura 3: Intuição do funcionamento do algoritmo t-SNE. Retirado de [3]

3 TRABALHOS RELACIONADOS

Alguns trabalhos recentes buscam tentar resolver a tarefa de classificar de forma automática discursos políticos em um espectro ideológico pré-definido. Normalmente as abordagens utilizam um dos dois tipos modelos de linguagens apresentados, em um contexto de aprendizado supervisionado.

A abordagem construída por [9] aplica uma *recursive neural network* (arquitetura de rede com estruturas de memória), que utiliza como entrada os vetores de representações distribuídas, para tentar identificar a posição política de sentenças em datasets conhecidos e em um dataset construído a partir de *crowdsourcing*. O espectro político considerado varia entre conservador, neutro e liberal. Os resultados apresentados são relativamente bons e é listado um conjunto de n-grams mais relacionados à cada espectro político.

Já [14] tenta aplicar modelos de linguagem probabilísticos (*Bayesian Hidden Markov Models*) para caracterizar sentenças de candidatos políticos em um espectro mais amplo, que vai desde a extrema esquerda até a extrema direita. Os modelos construídos baseiam-se em um corpus de sentenças que apresentam rótulos relacionados à esse espectro mais amplo.

O trabalho de [5] traz uma análise que foge à identificação de classes dentro de um espectro político. Diferentemente dos trabalhos anteriores, propõe a análise do conteúdo dos discursos políticos, categorizando-os em diferentes tópicos como economia, bem-estar, imigração, dentre outros. Utiliza uma *convolution neural network* para analisar os vetores das representações das palavras na sentença, classificando-a com a distribuição de probabilidade na saída da rede para cada um dos tópicos.

4 METODOLOGIA

O presente trabalho busca analisar, via modelos de linguagem neurais, o discurso proferido por deputadas e deputados de diferentes partidos políticos brasileiros, categorizados em um espectro político que vai desde a extrema esquerda, representada por partidos como PCdoB e PSOL até a direita/extrema direita, representada por partidos como PP, DEM e PSL. A categorização desse espectro político se baseia em [2] e [1]. Uma visão mais detalhada em relação às classificações de conservadorismo e progressismo pode ser observada na figura 4.

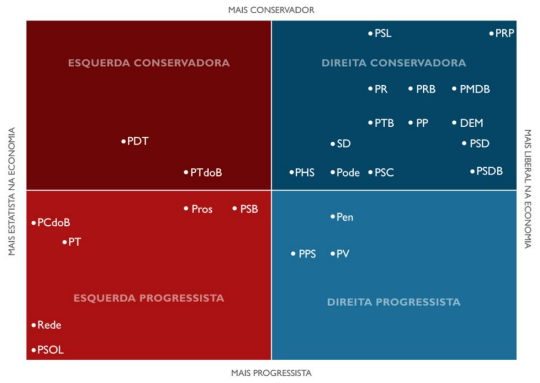

Figura 4: Detalhamento do espectro político dos partidos políticos do Brasil. Retirado de [1]

No meio desse espectro estão os partidos PT e PSDB, representando a esquerda/centro-esquerda e centro-direita, respectivamente. Tais partidos serão alvo das análises que serão elucidadas mais adiante.

4.1 Aquisição e pré-processamento dos dados

Os discursos das deputadas e deputados foram obtidos a partir de requisições feitas à API disponibilizada pela iniciativa Dados Abertos da Câmara dos Deputados. Foram coletados 314 discursos, proferidos no ano 2018 por 122 deputada(o)s de 21 partidos diferentes. Os discursos abrangem tópicos diversos que vão desde agradecimentos até votos sobre propostas e emendas políticas.

Para cada um dos discursos transformou-se a string original em tokens (split feito a partir do caractere de espaço) e em seguida utilizou-se funções de pré-processamento da biblioteca NLTK [6], como por exemplo a função para remoção de stopwords com base na língua Portuguesa, a fim de se extrair os tokens mais relevantes de cada discurso. Uma visualização do pré-processamento pode ser identificada na figura 5.

```
Nome do deputado: ORLANDO SILVA
>>> Discurso:
O SR. ORLANDO SILVA (PCdoB - SP. Pela ordem. Sem revisão do orador.) - Sr. Presidente, nós do PCdoB somos favorável à proposta apresentada, porque consideramos que estimular os taxistas a adquirir carros elétricos e carros híbridos impacta no meio ambiente e na sustentabilidade das nossas cidades. Oferecemos uma oportunidade para que esse serviço público tão importante para o Brasil, que é realizado pelos taxistas, se dê com veículos mais modernos e combustíveis do futuro. E desse modo o benefício final virá para a população.
O PCdoB vota favoravelmente à proposta.
>>> Tokens:
['sr', 'orlando', 'silva', 'pcdob', 'sp', 'ordem', 'revisao', 'orador', 'sr', 'presidente', 'pcdob', 'favoraveis'
, 'proposta', 'apresentada', 'porque', 'consideramos', 'estimular', 'taxistas', 'adquirir', 'carros', 'eletricos'
, 'carros', 'hibridos', 'impacta', 'meio', 'ambiente', 'cidades', 'oferecemos', 'oportunidade', 'servico', 'publi
co', 'tao', 'importante', 'brasil', 'realizado', 'taxistas', 'veiculos', 'modernos', 'combustiveis', 'futuro', 'd
esse', 'modo', 'beneficio', 'final', 'vira', 'populacao', 'pcdob', 'vota', 'favoravelmente', 'proposta']
```

Figura 5: Exemplo de pré-processamento básico feito para cada um discurso

4.2 Modelagem

Como mencionado anteriormente, foram utilizados modelos de linguagem neurais para realizar a tarefa de análise de discurso. No total, utilizou-se 4 tipos diferentes de modelos:

- `skip_s50`: Skip-Gram pré-treinado, utilizando vetores de dimensão 50
- `cbow_s50`: CBOW pré-treinado, utilizando vetores de dimensão 50
- `size_50_window_5_sg_1`: Skip-Gram treinado a partir do corpus dos discursos adquiridos, também utilizando vetores de dimensão 50
- `size_50_window_5_sg_0`: CBOW treinado a partir do corpus dos discursos adquiridos, também utilizando vetores de dimensão 50

Os modelos pré-treinados foram obtidos a partir do repositório de representações distribuídas vetoriais (Word Embeddings) do Núcleo Interinstitucional de Linguística Computacional (NILC) [7], do Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade Federal de São Paulo (USP).

A escolha de modelos pré-treinados em outros corpus (maiores e mais diversos, como no caso dos modelos do NILC) veio da necessidade de balancear os possíveis vieses captados ao se treinar os modelos somente no corpus dos discursos. Os resultados das análises serão estratificados por cada um dos modelos, a fim de se comparar o impacto da utilização de diferentes tipos de embeddings.

Uma vez aprendidos os embeddings para os modelos treinados no corpus dos discursos ou utilizando os embeddings pré-treinados é possível calcular a WMD entre dois discursos diferentes.

4.3 Análises

Duas macro análises são propostas na tentativa de se identificar similaridades e diferenças entre os discursos tomando por base o espectro político considerado no trabalho. Tomando por base os 2 partidos que governaram o país nos últimos 20 anos, PT (considerado de esquerda) e PSDB (considerado de direita), comparou-se o discurso da(o)s deputada(o)s desses partidos com o discurso da(o)s deputada(o)s dos partidos PCdoB, PDT, DEM e PP. Além disso, foi feita uma análise entre os discursos proferidos por diferentes deputada(o)s do PT e PSDB, tentando identificar um certo nível de coerência intra-partido.

Esses 4 últimos partidos (PCdoB, PDT, DEM e PP) foram escolhidos a fim de se obter representantes do espectro político considerado, indo desde a moderada/extrema esquerda com o PCdoB até a direita com o PP.

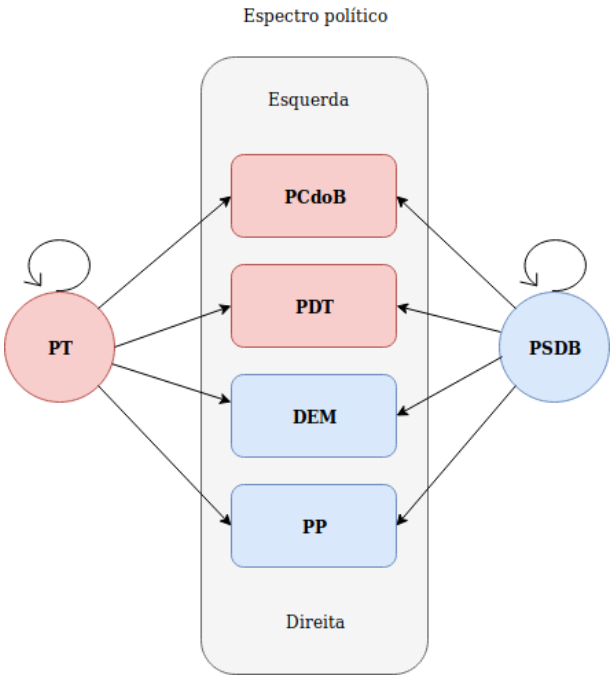

Figura 6: Visualização das comparações feitas entre partidos e entre deputados nas análises.

4.4 Análise de discursos entre partidos

Selecionou-se aleatoriamente 5 deputada(o)s de cada um dos partidos escolhidos, escolhendo 1 discurso proferido por cada um(a) delas(es). A seguir foi calculada a distância via WMD entre tais conjuntos de discursos como mostra a figura 6, em uma análise par-a-par.

O resultado é uma matriz quadrada ($5x5$) de distâncias onde cada célula indica a WMD entre um discurso do partido A e um discurso do partido B. Como foram utilizados 4 tipos de modelos de linguagem diferentes, teremos no total 4 matrizes de distância.

4.5 Análise de discursos entre deputados

Selecionou-se aleatoriamente 5 discursos da(o)s deputada(o)s que mais proferiram discursos em cada um dos partidos escolhidos. A seguir foi aplicada a mesma metodologia elucidada na explicação da análise anterior.

5 RESULTADOS E DISCUSSÃO

Se tirarmos a média da soma das distâncias das matrizes obtidas via análise explicada nas seções 4.3 e 4.4 e plotarmos tais distâncias em função de cada um dos 4 partidos do espectro, obtemos os seguintes resultados para PT e PSDB (incluindo os próprios PT e PSDB no eixo x, a fim de obter o resultado de uma análise intra-partido ou intra-deputado):

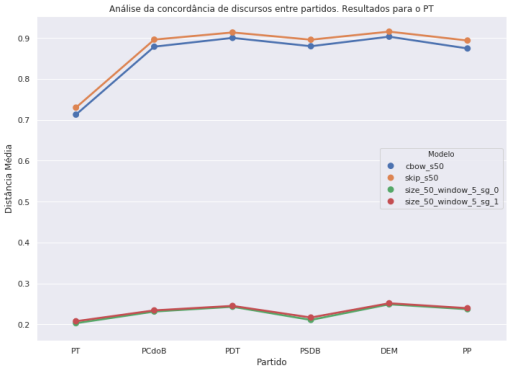

Figura 7: Análise da concordância de discursos entre partidos. Resultados para o PT.

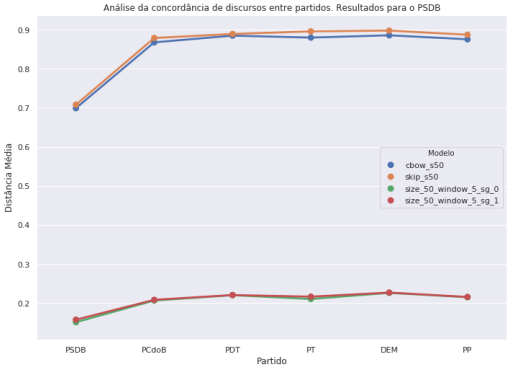

Figura 8: Análise da concordância de discursos entre partidos. Resultados para o PSDB.

Tomando por base a semelhança no espectro político, era esperado que discursos de deputada(o)s do PT se assemelhassem (ou seja, apresentassem distância média menor) mais aos discursos de deputada(o)s do PCdoB e PDT e menos aos discursos do DEM e do

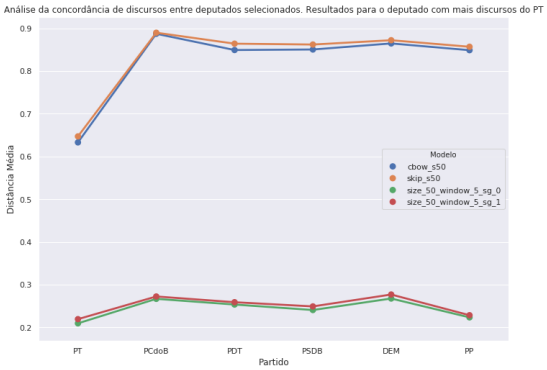

Figura 9: Análise da concordância de discursos entre deputados selecionados. Resultados para o deputado com mais discursos do PT.

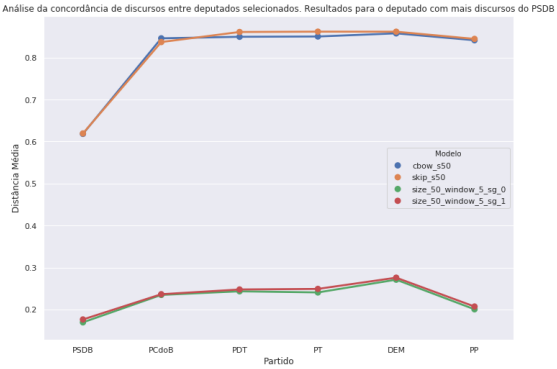

Figura 10: Análise da concordância de discursos entre deputados selecionados. Resultados para o deputado com mais discursos do PSDB

PP. Também era esperado que discursos de deputada(o)s do PSDB se assemelhassem mais aos discursos de deputada(o)s do DEM e PP e menos aos discursos do PCdoB e do PDT.

Observando os resultados obtidos não foi possível identificar a concordância esperada nas análises entre partidos e entre deputados. Algumas possíveis explicações para os resultados obtidos partem do número pequeno de discursos obtidos via API aberta da Câmara dos Deputados e da diversidade de assuntos presentes nos discursos.

Investigando com mais detalhes o conteúdo dos discursos obtidos foi possível identificar uma gama muito grande de tópicos diferentes tanto entre discursos de diferentes deputada(o)s como entre os discursos de uma mesma(o) deputada(o).

Com um número maior de discursos talvez os modelos de linguagem utilizados seriam mais capazes de identificar diferenças e semelhanças entre discursos, uma vez que haveriam mais sentenças representando os diferentes tópicos.

Ou seja, os resultados foram inconsistentes com os esperados possivelmente pelo fato de existirem poucos discursos com tópicos muito diferentes entre os partidos políticos. Assim, ao comparar poucos discursos de tópicos diferentes (representações distribuídas de palavras muito diferentes), não é possível estabelecer um certo limiar de concordância ou discordância entre eles.

Entretanto, os modelos de linguagem foram capazes de detectar uma certa concordância interna maior entre os discursos de um mesmo partido, como mostram as figuras 7 e 8, e entre os discursos de uma(um) mesma(o) deputada(o), como é possível ver nas figuras 9 e 10.

Um outro fato observado é em relação à magnitude das distâncias comparando-se os diferentes modelos de linguagem utilizados. Para os modelos pré-treinados em corpus maiores, foi possível identificar distâncias com magnitude maior em relação às distâncias obtidas ao se utilizar os modelos treinados no corpus de discursos.

Uma possível explicação para a diferença entre a magnitude das distâncias vem do fato do espaço de representações distribuídas dos modelos pré-treinados ser "muito mais preenchido"do que o espaço de representações dos modelos treinados com os discursos somente. Ou seja, o vocabulário dos modelos pré-treinados é muito maior que o vocabulário dos modelos treinados.

Assim, as relações semânticas construídas nesse espaço maior e mais preenchido podem ter mais detalhes que as relações construídas em um espaço menor e menos preenchido. Isso leva à situação em que pequenas mudanças nas dimensões dos vetores gerem cálculos de distância bastante diferentes. Desse modo, ao calcular a WMD entre os discursos utilizando os diferentes modelos obtêm-se magnitudes diferentes.

Por fim, utilizou-se o algoritmo t-SNE para visualizar o vetor médio dos embeddings gerados pelo modelo `size_50_window_5_sg_1` das TOP 25 palavras mais presentes em todos os discursos de cada partido a fim de também tentar identificar algum tipo de relação de proximidade tomando por base o espectro político.

Novamente os resultados observados na figura 11 não refletiram o esperado possivelmente pelas mesmas explicações dadas anteriormente.

6 CONCLUSÕES

A partir da execução do trabalho foi possível aplicar os conceitos relacionados à modelos de linguagem em um cenário real de comparação de discursos de deputada(o)s provenientes de diferentes partidos políticos brasileiros, que se encaixam em um espectro político esquerda-direita.

Devido à possível baixa qualidade dos dados adquiridos (em termos de diversidade e quantidade), os resultados encontrados não refletiram à concordância esperada entre os discursos de diferentes partidos desse espectro político. Entretanto, os modelos de linguagem foram capazes de identificar uma certa concordância de discursos intra-partido e intra-deputada(o).

Também foi possível identificar uma diferença de cálculo de distância (via WMD) entre os modelos de linguagem pré-treinados e os modelos treinados no corpus de discursos adquiridos.

Possibilidades para trabalhos futuros giram em torno da aquisição de dados de maior qualidade bem como a análise de outros tipos de configuração de modelos de linguagem neurais.

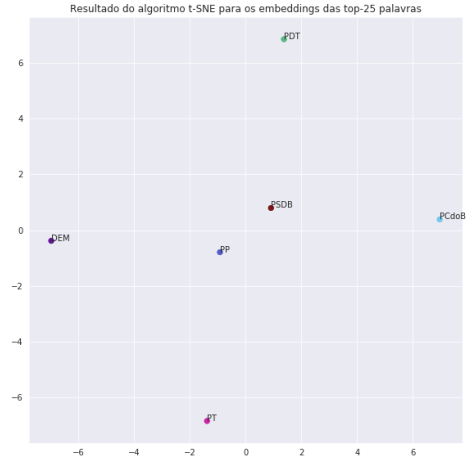

Figura 11: t-SNE aplicado nas TOP 25 palavras presentes nos discursos de cada partido analisado.

REFERÊNCIAS

[1] [n. d.]. Espectro político dos partidos do Brasil. https://www.bbc.com/portuguese/brasil-41058120.

[2] [n. d.]. Lista de partidos políticos do Brasil. https://pt.wikipedia.org/wiki/Lista_de_partidos_políticos_do_Brasil.

[3] [n. d.]. tSNE visually explained. http://blog.thegrandlocus.com/2018/08/a-tutorial-on-t-sne-1.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[5] Aritz Bilbao-Jayo and Aitor Almeida. 2018. Political discourse classification in social networks using context sensitive convolutional neural networks. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. 76–85.

[6] Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 31.

[7] Erick R Fonseca, João Luís G Rosa, and Sandra Maria Aluísio. 2015. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. *Journal of the Brazilian Computer Society* 21, 1 (2015), 2.

[8] Anna Huang. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 49–56.

[9] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1113–1122.

[10] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*. 957–966.

[11] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[14] Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 91–101.

[15] Andreas Stolcke. 1994. *Bayesian learning of probabilistic language models*. Ph.D. Dissertation. University of California, Berkeley.

[16] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. 2009. Dimensionality reduction: a comparative. *J Mach Learn Res* 10 (2009), 66–71.