Análise da Evolução dos Discursos de Pré-candidatos à Presidente por meio de Representações Linguísticas Vetoriais[1]

Kid Valeriano, Aline Paes e Daniel de Oliveira

Instituto de Computação, Universidade Federal Fluminense - Brasil
kvaleriano@id.uff.br, {alinepaes,danielcmo}@ic.uff.br

Abstract. Comumente, os pré-candidatos aos cargos governamentais expressam suas opiniões e plataformas de campanha em discursos informais, previamente ao período oficial. Esse comportamento é essencial para que o eleitor conheça as ideologias e plataformas de campanha, de forma a tomar sua decisão de voto. No processo decisório, o eleitor pode considerar a semelhança entre discursos de diferentes candidatos, como o discurso varia ao longo do tempo, e qual a adequação do discurso aos temas mais relevantes para a sociedade. Entretanto, analisar e capturar tais aspectos a partir dos discursos informais é uma tarefa difícil para o eleitor, dado o volume de informação disponibilizada por diversos veículos de comunicação, e o viés político de alguns deles. Assim, nesse artigo, propomos uma ferramenta de análise de discurso político baseada em técnicas de Aprendizado de Representações Linguísticas para auxiliar o eleitor na sua decisão. Resultados obtidos a partir dos discursos dos pré-candidatos ao cargo de Presidente do Brasil em 2018 permitem verificar como os candidatos se comportam em termos de seus próprios discursos e dos discursos de seus concorrentes.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Machine Learning Applications; I.2.7 [**Natural language processing**]: Discourse

Keywords: doc2vec, natural language processing, discourse analysis

1. INTRODUÇÃO

Durante o período de campanha eleitoral, é esperado que os candidatos a cargos governamentais apresentem suas plataformas de governo para a população. A campanha é disseminada tanto a partir de meios tradicionais, como TV, rádio, e mídia impressa, como também, mais recentemente, a partir das mídias digitais. Usualmente, esse período de campanha é limitado a poucos meses antes do dia da votação. Nas eleições Brasileiras de 2018, por exemplo, esse período será de 45 dias[2]. A partir dessa data, todos aqueles oficialmente inscritos como candidatos devem seguir uma série de regras que regulamentam a propaganda eleitoral, incluindo o uso de linguagens de sinais e propaganda apenas em páginas *web* com a terminação *.can.br*[3].

Uma das formas mais abrangentes de propaganda é a veiculada gratuitamente em emissoras abertas de TV e rádio[4]. Porém, nas eleições Brasileiras, o tempo de propaganda eleitoral gratuita depende das coligações estabelecidas entre os partidos, o que faz com que alguns candidatos possuam muito tempo de propaganda gratuita, enquanto outros conseguem apenas poucos segundos. De qualquer forma, nem todos os eleitores dispõem de tempo ou interesse para assistir todas as propagandas gratuitas.

[1]Os autores gostariam de agradecer a CAPES, CNPq e FAPERJ pelo apoio financeiro.
[2]http://www.justicaeleitoral.jus.br
[3]https://www.cgi.br/resolucoes/documento/2008/008
[4]https://exame.abril.com.br/brasil/como-a-propaganda-eleitoral-afetou-as-ultimas-eleicoes-presidenciais/

Assim, é cada vez mais comum que os pré-candidatos comecem a disseminar suas ideias e planos de governo antes do período oficial de propaganda eleitoral, utilizando mídias digitais e redes sociais, *e.g.* Facebook, Twitter, etc. De forma similar, a imprensa também entrevista e publica conteúdo relacionado aos pré-candidatos antes do período oficial. Em ambos os casos, o eleitor pode dar início ao seu processo de decisão de voto o quanto antes, usando, para tanto, as informações obtidas a partir de tais entrevistas, relatos, e vídeos disponibilizados pelos candidatos.

Embora essa prática seja útil para dar mais tempo para o cidadão tomar uma decisão consciente e bem fundamentada, o volume de informações pode ser tão grande que ele terá dificuldades de processá-las. Logo, aspectos relevantes como: (i.) a ideologia e características políticas do candidato, (ii.) seu posicionamento frente às questões consideradas relevantes pelo cidadão, (iii.) como o candidato defende suas ideias ao longo do tempo (*i.e.,* se o discurso do candidato é consistente), (iv.) o que o candidato tem a dizer a partir de temas atuais e pontuais, (v.) e como ele se compara em relação aos demais candidatos, podem passar despercebidas ou não serem passíveis de análise pelo eleitor.

Assim, nesse artigo, propomos uma ferramenta direcionada a auxiliar o eleitor no processo de tomada de decisão. A ferramenta tem dois objetivos principais: (i.) comparar a similaridade dos discursos de um mesmo candidato ao longo do tempo e (ii.) comparar a similaridade dos discursos de dois candidatos. Como as comparações devem ser feitas de forma automática, a partir de conteúdo linguístico, a ferramenta se baseia em Aprendizado de Máquina para aprender uma representação linguística latente [Le and Mikolov 2014] dos discursos. Especificamente, utilizamos a técnica de vetorização automática não-supervisionada de documentos, denominada de *Doc2Vec* [Le and Mikolov 2014; Dai et al. 2015]. Dados os documentos representados em formato vetorial, é possível utilizar medidas de distância para verificar aqueles que estão mais próximos uns dos outros, como acontece com a já consolidada técnica Word2Vec [Mikolov et al. 2013; Mikolov et al. 2013]. De forma a mostrar a utilidade da ferramenta desenvolvida, utilizamos como estudo de caso vídeos e entrevistas dos pré-candidatos ao cargo de Presidente do Brasil na eleição de 2018. Os conteúdos textuais referentes ao que os candidatos falaram são extraídos automaticamente a partir dos áudios.

Este artigo se encontra organizado em 4 seções além da introdução. A Seção 2 discute o *Doc2Vec*. A Seção 3 apresenta como computar similaridade dos discursos com *Doc2Vec*. A Seção 4 discute os trabalhos relacionados, e, finalmente, a Seção 5 conclui o artigo.

2. APRENDIZADO DE REPRESENTAÇÕES LINGUÍSTICAS EM DOCUMENTOS

Até alguns anos atrás, as técnicas de *bag of words* (BOW) e *Bag-of-n-grams* [Harris 1954; Zhang et al. 2010] eram as mais utilizadas para transformar textos em um conjunto de atributos, de forma a aplicar técnicas de Aprendizado de Máquina para a indução de padrões. Rudemente falando, o método BOW consiste em transformar o texto em um conjunto de *tokens*, considerar que tais *tokens* são os atributos da tarefa de aprendizado, e utilizar a frequência do *token* (ou o inverso da frequência) em um exemplo como valor associado aos atributos. Porém, por se basear em contagem, esse método falha em situações em que o padrão a ser extraído deve levar em consideração aspectos semânticos.

Uma das causas é que os aspectos semânticos das palavras em um texto podem variar de acordo com o contexto considerado. Assim, supondo, por exemplo, que os valores associados a um atributo podem variar entre 0 e 1, os *tokens* "rei"e "rainha"em um contexto de realeza deveriam ter valores próximos um ao outro, e próximos a 1. Por outro lado, em um contexto de gênero, os valores associados a essas mesmas duas palavras deveriam ser distantes um do outro , uma vez que tratam de gêneros diferentes. Já em um contexto alimentício, essas palavras deveriam ter valores muito baixos, embora próximos um do outro. Ou seja, definir um conjunto de atributos que generalize sobre diversos contextos, e atribuir valores apropriados para tais atributos, é uma tarefa difícil de ser realizada manualmente e propensa a erros, principalmente devido à subjetividade.

Para contornar esse problema, tem se tornado uma prática comum utilizar representações vetoriais

numéricas para associar valor aos componentes de um texto (*e.g.,* suas palavras). Tais representações são mais conhecidas como *embeddings* [Collobert et al. 2011], e normalmente são definidas como vetores de 300 dimensões, aprendidos de forma automática a partir de diversos textos. Assim, cada dimensão pode refletir um contexto distinto, e o valor associado à dimensão é aprendido de acordo. Espera-se que, ao final do processo de aprendizado, as palavras com semânticas mais próximas sejam mapeadas para posições mais próximas no espaço vetorial. As implementações mais utilizadas de tais técnicas, *word2vec* [Mikolov et al. 2013] – que por sua vez implementa os algoritmos *Skip-gram* [Guthrie et al. 2006] e *CBOW* [Mikolov et al. 2013] – e GloVe [Pennington et al. 2014], utilizam redes neurais com uma camada escondida. Os vetores referentes aos atributos de palavras são extraídos a partir dos pesos da camada oculta, fazendo com que essa forma de aprendizado receba o nome de modelos neurais de linguagem [Bengio et al. 2003]. Brevemente falando, o objetivo do modelo é maximizar o valor de:

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p\left(w_t|w_{t-k},\ldots,w_{t+k}\right) \tag{1}$$

Onde w_i representa uma palavra em uma sequência de palavras $w_1, w_2, \ldots, w_T$, e $w_{t-k}, \ldots, w_{t+k}$ representa uma janela de palavras de tamanho t, onde $w_{t-k}, w_k, w_{t+k} \subset w_1, w_2, \ldots, w_T$. Cada tarefa de predição é usualmente definida como um classificador *softmax*, como a seguir:

$$p\left(w_t|w_{t-k},\ldots,w_{t+k}\right) = \frac{e^{y_{wt}}}{\sum_i e^{y_i}} \tag{2}$$

Onde y_i é o logaritmo não normalizado da probabilidade da palavra i ser a saída do modelo, computado como:

$$y = b + Uh\left(w_{t-k},\ldots,wt+k;W\right) \tag{3}$$

em que U e b são os pesos do classificador e h é ou a concatenação ou a média dos vetores de palavras em W. Os modelos neurais de linguagem são treinados com gradiente descendente, onde o gradiente é obtido a partir do algoritmo de retro-propagação [Rumelhart et al. 1986].

Como o objetivo do presente artigo é detectar similaridade entre documentos, o ideal é que estes possam ser dispostos diretamente em um espaço vetorial, da mesma forma que as palavras. Assim, torna-se possível verificar aqueles que se encontram mais próximos, seguindo uma métrica de distância de vetores, e classificá-los como mais semanticamente similares. Para tanto, nos beneficiamos do modelo *Doc2Vec* [Le and Mikolov 2014], que tem como principal funcionalidade criar representações vetoriais para fragmentos de textos, independente de seus tamanhos. Tal método se baseia nos mesmos modelos de aprendizado de representações vetoriais de palavras, mas, além da matriz de vetores de palavras W, uma matriz de vetores de documentos D também é treinada. Assim, a equação 2 é reescrita como a seguir:

$$y = b + Uh\left(w_{t-k},\ldots,wt+k;W,D\right) \tag{4}$$

Duas implementações do Doc2Vec são mais usadas: $PD-DVM$, derivada do método *Skip-gram* e $PD-DBOW$, derivado do CBOW. O modelo $PV-DBOW$, em particular, recebe como entrada a matriz de documentos e devolve como saída palavras que estão associadas ao documento. Nesse caso, o vetor $d_i \in D$ associado ao documento pode ser visto como uma nova palavra, que será compartilhado entre todos os contextos oriundos do mesmo documento, mas não entre todos os documentos. A matriz de vetores de palavras W, por outro lado, é compartilhada entre todos os documentos.

Em tempo de inferência, na presença de um novo documento $doc_k \notin Docs$, onde $Docs$ é o conjunto de documentos usados para o treinamento, é necessário executar o método do gradiente descendente para obter o vetor representativo de doc_k. Para tanto, uma nova coluna é adicionada a D e os vetores em D são ajustados seguindo o gradiente, mas mantendo U, W e b fixos. Para verificar se dois

documentos são similares, comumente usa-se uma medida de distância entre os respectivos vetores, como por exemplo, a similaridade de cosseno. A Equação 5 exibe o cálculo de tal medida, onde **A** e **B** são vetores e A_i e B_i são seus componentes, respectivamente.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} = \frac{\sum_{i=0^n} A_i B_i}{\sqrt{\sum_{i=0^n} A_i^2} \sqrt{\sum_{i=0^n} B_i^2}} \tag{5}$$

3. COMPUTANDO A SIMILARIDADE DE DISCURSOS DE PRÉ-CANDIDATOS COM O DOC2VEC

A ferramenta desenvolvida nesse artigo consiste de quatro componentes principais: (i.) coleta de dados; (ii.) pré-processamento de dados; (iii.) treinamento do modelo com Doc2Vec[5]; e (iv.) detecção de similaridade a partir do modelo treinado. Em seguida, são gerados resultados que mostram o comportamento de similaridade entre os discursos dos candidatos.

3.1 Componentes do Processo Experimental

Coleta e organização de dados. Para a coleta de dados dos discursos dos pré-candidatos, utilizamos os vídeos disponibilizados abertamente na plataforma YouTube®. A seguir, o conteúdo falado no vídeo é extraído em formato textual, usando as ferramentas *Downsub*[6], que extrai o áudio em formato de legenda, e *Subtitletools* [7], que converte arquivos de legenda em texto plano. Os textos são, então, organizados por candidatos e por data de divulgação que nós chamamos janelas de tempo compreendendo 'Anteriores' ao mês de Outubro, 'Outubro','Novembro', até o mês 'Junho'. Foi adotado como critério de seleção vídeos de diferente duração desde 2 minutos até 2 horas aproximadamente, dos pré-candidatos à presidência do Brasil nas eleições de 2018. No total, foram coletados 353 discursos.

Pré-processamento dos textos. Apenas quatro passos de tratamento fazem parte da ferramenta desenvolvida: (i.) conversão para letras minúsculas, (ii.) extração de palavras a partir dos textos (tokenização), (iii.) remoção de palavras sem significado semântico atrelado (remoção de *stop-words*, incluindo números), e (iv.) *stemização* de Porter, que é a transformação de uma palavra para a sua base ou raiz, ao remover de forma heurística os sufixos. Cada um desses passos de pré-processamento foram executados usando a biblioteca NLTK [Bird and Loper 2004]. Por usar o Doc2Vec, não é usual que sejam feitos tratamentos adicionais, que poderiam atrapalhar a captura da semântica dos documentos.

Treinamento dos vetores de documentos. Para alimentar a entrada do Doc2Vec, a ferramenta pode seguir três caminhos: (i.) conversão para minúsuculas e tokenização, (ii.) seguido de remoção de *stop-words* e *stemização*, ou (iii.) seguido de remoção de *stop-words* apenas. Os demais parâmetros do Doc2Vec seguem a recomendação em [Lau and Baldwin 2016], o uso da implementação PV-DBOW, incluindo 300 como o tamanho dos vetores de representação, 15 como o tamanho da janela de contexto, 200 épocas de treinamento, entre outros.

3.2 Resultados Experimentais

Os resultados apresentados neste artigo são a interpretação do uso de vetores resultantes do modelo Doc2vec aplicado aos discursos políticos, usando a medida de similaridade de cosseno [Dai et al. 2015],

[5] `https://radimrehurek.com/gensim/models/doc2vec.html`
[6] `https://downsub.com/`
[7] `https://subtitletools.com`

computada no espaço vetorial dos documentos representados como vetores. Selecionamos os pré-candidatos que haviam manifestado a intenção de se candidatar, para mostrar os resultados obtidos nesta pesquisa, por razões de espaço serão analisados os candidatos: Jair Bolsonaro, Ciro Gomes, Geraldo Alckmin, Manuela D'Avila e Marina Silva.

Semelhança mútua. A semelhança mútua é calculada encontrando a semelhança entre dois pares de discursos, onde ambos os documentos são mais semelhantes um ao outro. Ou seja, suponha um documento A cujo documento mais semelhante a ele seja B, e, da mesma forma, o documento B possui maior semelhança com o documento A. Ao calcular a semelhança mútua, espera-se que os dois pares de discursos contenham uma forte semelhança semântica em nosso espaço vetorial, se a semelhança fora alta poderia tratar-se de discursos parecidos usando a mesmas frases ou palavras. O resultado da análise de discursos que apresentam similaridade mútua apresentado na Figura 1.

A semelhança mútua entre discursos do mesmo candidato de cor azul: observou-se que a maioria dos candidatos tem uma maneira de pensar sobre um tema específico, sendo constante ou repetitiva ao longo de sua campanha, *e.g.,* (a) Manuela D'Avila: propostas para valorização do trabalho e (b) Manuela D'Avila: 1 de Maio dia do trabalho. A semelhança mútua entre discursos de candidatos diferentes de cor laranja: representa que ambos os candidatos possuem ideais semelhantes ou que se tenta replicar parte de algum pensamento ou propostas de outro candidato, *e.g.,* (a) Sabatina de Pré-Candidatos a Presidência da República com Ciro Gomes e (b) Sabatina de Pré-Candidatos a Presidência da República com Marina Silva. Em ambos os discursos os candidatos foram arguidos sobre a economia (impostos e auditorias fiscais).

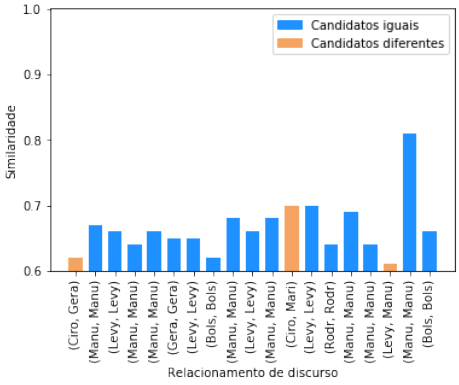

Fig. 1: Semelhança mútua entre pares de discursos de pré-candidatos a Presidente do Brasil

Semelhança média entre candidatos. Para computar a semelhança entre mais de um discurso, utilizamos os vetores dos documentos para calcular a semelhança média, conforme apresentado na Equação 6. A similaridade entre os candidatos é calculada usando a média de todas as semelhanças obtidas por pares de dois vetores pertencentes a candidatos distintos. Assim, cada discurso feito por um determinado candidato, em uma determinado janela de tempo, é comparado com os discursos de outro candidato, obtendo como resultado a similaridade média de cada candidato em relação aos demais.

$$SMC\_D2V = \frac{1}{N > 0} \sum_{j=2}^{N} \sum_{i=1}^{j-1} similarity(dv_j, dv_i) \tag{6}$$

A Tabela I exibe, para cada candidato, o candidato com o qual ele obteve a maior similaridade em seus discursos, *i.e.,* similaridade máxima. A similaridade máxima pode indicar que eles possuem linhas

Tabela I: Similaridade Máximaentre os discursos dos pré-candidatos a Presidente do Brasil

Candidatos	**Similaridade Máxima**
Jair Bolsonaro	0.353 (Manuela)
Ciro Gomes	0.344 (Geraldo)
Geraldo Alckmin	0.37 (Manuela)
Manuela D'Avila	0.37 (Geraldo)
Marina Silva	0.354 (Manuela)

Fig. 2: Candidato mais constante: tamanho 3 de *Manuela D'Avila*

ideológicas ou propostas semelhantes, ou ainda que eles podem abordar assuntos parecidos em seus respectivos discursos, como o uso de tópicos similares nos discursos, mesmo que tratados com linhas de pensamento antagônicas, como também propostas e ideologias distintas relativas aos mesmos tópicos. Na Tabela I podemos perceber que o pré-candidato Jair Bolsonaro possui maior similaridade média em relação a candidata Manuela D'Avila. Apesar dos pré-candidatos possuírem linhas ideológicas distintas, os dois discutem sobre temas semelhantes em seus discursos como movimento feminino e LGBT, cada qual com uma perspectiva diferente.

Discursos constantes de candidatos. O uso das janelas do tempo definidas anteriormente assume um significado vital nesta métrica. A ideia é verificar quais candidatos mantém o discurso constante por uma maior quantidade de janelas de tempo, ao longo da campanha. Para o cálculo tomam-se inicialmente todos os vetores dos discursos mais antigos do candidato em avaliação, de forma que possamos calcular os dez discursos mais similares ao mais antigo, e repetindo esse processo partindo dos discursos mais antigos até chegar à atualidade. Nesse cálculo, utilizamos uma similaridade mínima de 0,5, para que pelo menos tenhamos 50% de semelhança entre os discursos. Além disso, utilizamos uma constante de tempo 1 por janela, *i.e.*, se considerarmos os discursos de um candidato de outubro a junho, calculamos iterativamente o mais semelhante em relação a Novembro, depois em relação a Dezembro, até Junho. A Fig. 2 exibe o grafo de continuidade da candidata que se manteve constante por mais tempo, Manuela D'Avila. Ser constante pode nos levar a pensar que o candidato tem seus ideais bem definidos, sem mudar seu modo de pensar, mesmo que existam fatores que o estimulem a fazê-lo durante toda a sua campanha, ou também faça seus discursos na mesma direção durante períodos consecutivos, sem apresentar, no entanto, uma evolução natural em seus pensamentos. Conforme apresentado na Tabela III, o pré-candidato Geraldo Alckmin não possui discursos constantes (muda de assunto discursado constantemente), mas os candidatos Jair Bolsonaro, Ciro Gomes e Marina Silva têm como a maior tava de constância 2, Manuela D'Avila, por sua vez, tem muitos caminhos constantes com valor 2, que ademais têm grandes semelhanças entre os discursos constantes, destacando que é o candidata mais constante no que se refere aos temas discursados.

Coerência de discursos por período de tempo. Chamamos de coerência à proximidade semântica que tem os discursos proferidos pelos candidatos durante os períodos de campanha. A avaliação foi realizada usando a similaridade entre os discursos dos candidatos durante uma janela de tempo, *i.e.* os discursos de um mesmo candidato foram comparados variando-se a data em que o mesmo foi realizado. Os resultados observados na Figura 3 mostram a coerência dos diferentes discursos realizados por um determinado candidato em uma janela de tempo. De acordo com a Fig. 3, o

Tabela II: Tres caminhos constantes por candidato

Candidato	Caminho	Arestas	Tamanho	Similaridade
Bolsonaro	Maio ->Junho	4	1	0,52 / 0,51 / 0,56 / 0,57
Bolsonaro	Marco ->Abril ->Maio	2	2	0,55 / 0,58
Bolsonaro	Maio ->Junho	1	1	0,58
Ciro	Abril ->Maio ->Junho	3	2	0,54 / 0,6 / 0,59
Ciro	Maio ->Junho	2	1	0,55 / 0,58
Gera	Maio ->Junho	2	1	0,52 / 0,57
Manuela	Março ->Abril ->Maio	2	2	0,68 / 0,5
Manuela	Março ->Abril ->Maio ->Junho	3	**3**	0,55 / 0,69 / 0,61
Manuela	Março ->Abril ->Maio	2	2	0,62 / 0,6
Marina	Maio ->Junho	2	1	0,55 / 0,56

candidato Geraldo Alckmin na janela anterior a Outubro de 2017 - "ant"na Fig. 3 atinge a maior coerência, enquanto que Ciro Gomes apresenta a menor coerência. Tanto Ciro quanto Alckmin foram os que mais apresentaram variações de coerência ao longo do tempo, enquanto que os candidatos Marina Silva, Manuela D'Avila e Jair Bolsonaro foram coerentes ao longo de suas campanhas. Porém, esse resultado pode refletir que os 3 candidatos sempre tratam dos mesmos temas ou que eles de fato mantém um discurso semanticamente similar ao longo do tempo.

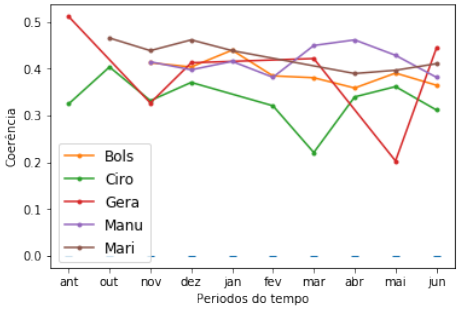

Fig. 3: Coerência de discursos por período

4. TRABALHOS RELACIONADOS

Embora no melhor do nosso conhecimento não existam trabalhos similares ao aqui exposto no contexto de eleições brasileiras, trabalhos anteriores já tratam do tema de similaridade de discursos políticos. Em [Greene and Cross 2017] o objetivo principal era extrair o tópico mais relevante utilizando a decomposição não negativa da matriz de fatorização de textos (NMF). O processo foi realizado em duas camadas, onde, ao aplicar o NMF na primeira camada obtém-se o tópico principal de cada período de tempo, e ao aplicar o NMF em uma segunda camada, a partir dos resultados da primeira, foram capturados temas de recorrência da Agenda Política do Parlamento Europeu. Esta abordagem, assim como a nossa, é não-supervisionada. Em [Gautrais et al. 2017] os autores também usaram um modelo de duas camadas. Porém, enquanto a primeira camada também objetivou extrair o tema do texto, na segunda camada, usando um algoritmo chamado *Signature Model*, foi extraído o tópico recorrente ao longo do tempo, com o uso de programação linear. Como se observa nesses dois trabalhos, a análise da evolução dos candidatos é baseada somente nos temas. Em [Azarbonyad et al. 2017] foi proposta uma abordagem para detectar mudanças semânticas entre diferentes pontos de vista aplicados aos discursos políticos, tendo como base o vetor de palavras computado com o Word2Vec. Basicamente, computa-se a distância entre uma palavra em um espaço e a mesma palavra em outro espaço, para verificar se houve mudança no significado de ambas.

5. CONCLUSÕES

Neste artigo, apresentamos alguns resultados práticos utilizando a representação vetorial de discursos informais curtos de pré-candidatos ao cargo de Presidente do Brasil. Para obter a representação no formato de vetores multidimensionais, nos beneficiamos da abordagem Doc2Vec, que, da mesma forma que o Word2Vec faz com palavras, tenta extrair a semântica dos documentos a partir do processo de aprendizado. Com o Doc2Vec e a medida de distância de cosseno, foi possível exemplificar como analisar a evolução dos candidatos ao longo da campanha política. Pelos resultados, foi possível observar que nem sempre conseguimos medir a similaridade semântica considerando políticas e ideologias, uma vez que não levamos em consideração os temas tratados nos discursos. Como trabalho futuro, pretendemos avaliar a similaridade entre ideologias políticas de diferentes candidatos, no que diz respeito a temas específicos discutidos por eles, bem como análises mais extensas conectando os candidatos aos seus tópicos de discurso, e outras formas de computar distância entre os vetores.

REFERENCES

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., and Kamps, J. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, pp. 1509–1518, 2017.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research* 3 (Feb): 1137–1155, 2003.

Bird, S. and Loper, E. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, pp. 31, 2004.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12 (Aug): 2493–2537, 2011.

Dai, A. M., Olah, C., and Le, Q. V. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*, 2015.

Gautrais, C., Cellier, P., Quiniou, R., and Termier, A. Topic signatures in political campaign speeches. In *EMNLP 2017-Conference on Empirical Methods in Natural Language Processing*, 2017.

Greene, D. and Cross, J. P. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis* 25 (1): 77–94, 2017.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*. pp. 1–4, 2006.

Harris, Z. S. Distributional structure. *Word* 10 (2-3): 146–162, 1954.

Lau, J. H. and Baldwin, T. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. pp. 78–86, 2016.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. pp. 1188–1196, 2014.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pp. 3111–3119, 2013.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 746–751, 2013.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pp. 1532–1543, 2014.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature* 323 (6088): 533, 1986.

Zhang, Y., Jin, R., and Zhou, Z.-H. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* 1 (1-4): 43–52, 2010.