https://eventos.utfpr.edu.br//sicite/sicite2020

Mineração de texto e aprendizado de máquina na identificação de posicionamento em discursos: um estudo de caso

Text mining and machine learning on identification of positioning in speeches: a case study

Vinícius Couto Tasso
vintas@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil

André Eugenio Lazzaretti
lazzaratti@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil

RESUMO

Processamento de linguagem natural é um ramo da inteligência computacional que permite aos computadores entender a linguagem humana. É um tópico amplamente estudado na literatura, e suas aplicações são altamente difundidas atualmente. Este trabalho apresenta um estudo de caso da aplicação de métodos computacionais de mineração de texto na descoberta de conhecimento em bases de dados. O objetivo é encontrar padrões para identificar anomalias em pronunciamentos de deputados federais sobre um tema em comum. A base de dados utilizada é composta por discursos de parlamentares sobre a Reforma da Previdência (PEC 06/19). Os discursos foram proferidos entre o início do ano de 2019 e 10 de julho, data da votação da proposta em primeiro turno. Os resultados sugerem que o método proposto é robusto e promissor para a aplicação apresentada, alcançando um valor de F1 de 0.99 no subconjunto de teste.

PALAVRAS-CHAVE: Mineração de dados. Processamento de Linguagem Natural. Aprendizado do computador.

ABSTRACT

Natural language processing is a branch of computational intelligence that allows computers to understand human language. The topic is discussed widely in the literature, and its applications are widespread today. This work presents a case study of the application of computational methods of text mining in the discovery of knowledge in databases. The goal is to find patterns to identify anomalies in statements by federal deputies on a common theme. The database used is composed of a collection of speeches by parliamentarians on the Brazilian Social Security Reform. Speeches date from the beginning of 2019 to July 10th. The results suggest that the proposed method is robust and promising for the presented application, reaching a F1 Score of 0.99 on the test subset.

KEYWORDS: Data mining. Natural language processing. Machine learning.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

INTRODUÇÃO

A Reforma da Previdência foi uma Proposta de Emenda à Constituição amplamente debatida durante o decorrer do ano de 2019, propondo alterações no sistema de Previdência Social brasileiro. O projeto tinha, como uma de suas principais motivações, promover um caminho confiável e eficiente para geração de superávit primário (ROBERTO, 2019, p.25). A polêmica que envolveu os debates e discussões pode ser explicada pela dualidade estabelecida por opiniões favoráveis e contrárias à proposta, portanto divergentes entre si.

Este estudo se propõe a realizar uma análise exploratória nos discursos coletados, com o objetivo de encontrar anomalias e incoerências. Essas anomalias são definidas por dados que sugerem fazer parte de um grupo (e.g. favorável), mas que pertencem a outro (e.g. contrário), sem motivo aparente. Tais ocorrências podem indicar incoerências ou mudanças de posicionamento por parte de um parlamentar, por exemplo. Não há conhecimento prévio sobre o conjunto de dados utilizados, mas pressupõe-se que essas incoerências existam em alguma quantidade.

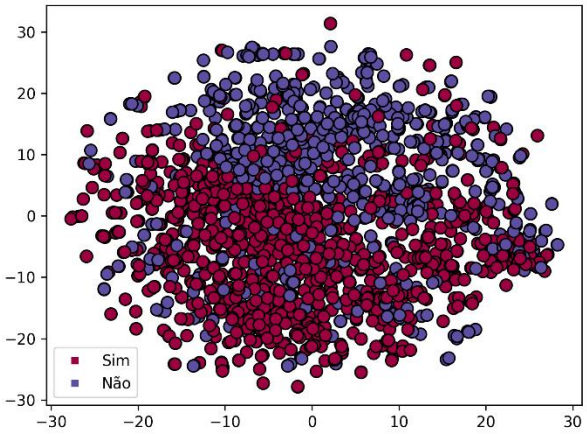
MATERIAL E MÉTODOS

Os dados utilizados para a realização deste trabalho são provenientes da plataforma de Dados Abertos da Câmara dos Deputados. O banco de dados é composto por pronunciamentos que tenham sido registrados nos sistemas da Câmara com palavras-chave remetentes ao tema de interesse. Os discursos foram anonimizados para evitar que metadados (e.g. nome do orador e filiação partidária) interferissem nos resultados. O banco de dados obtido é desbalanceado e constituído por duas classes distintas (Sim e Não) que refletem o voto em primeiro turno do orador.

O conjunto é composto por 1872 discursos únicos, dos quais a classe de pronunciamentos favoráveis representa cerca de 40%. Portanto os discursos contrários à proposta compuseram os 60% restantes. Para a realização dos experimentos, o conjunto de dados foi dividido em um subconjunto de treino, composto por 80% da totalidade dos dados, e um subconjunto de teste com o restante. Apenas o conjunto de teste passou por uma avaliação manual das classes atribuídas aos dados.

A Figura 1 apresenta uma visualização de todo o conjunto de dados (representado conforme a seção Representação indica) com dimensionalidade reduzida através do algoritmo *T-distributed Stochastic Neighbor Embedding* (t-SNE). A visualização em duas dimensões do espaço de *features*, que originalmente tem milhares de dimensões, sugere que os dados não são distribuídos no espaço de maneira regular e homogênea, e que se sobrepõem com frequência.

Figura 1 – Visualização em 2D dos dados com t-SNE



Fonte: Autoria própria.

PRÉ-PROCESSAMENTO

Para reduzir ruídos e garantir maior confiabilidade dos resultados obtidos, técnicas de tratamento e limpeza de dados foram empregadas, procedimento crucial em processamento de linguagem natural (KUMAR, 2019, p. 498).

Intromissões na fala do orador, anotações do transcritor e documentos anexados na transcrição foram considerados ruído e removidos. Após a remoção de ruídos externos, foi realizada a remoção de ruídos inerentes aos próprios dados de interesse: as palavras consideradas pouco informativas (e.g. preposições e conjunções), denominadas de *stop words*. As *stop words* são comuns e abundantes, e apesar de sua importância sintática para a formação de sentenças, carregam um valor semântico muito baixo (RAULJI; SAINI, 2016, p. 15-16). Portanto, foram filtradas e desconsideradas.

Com o intuito de preservar parte do contexto no qual as palavras estão inseridas, o procedimento de tokenização foi utilizado. Nessa representação, o texto é dividido em combinações de até n *tokens* (n-gramas) vizinhos entre si, resultando em tuplas de palavras relacionadas entre si.

Por fim, objetivando maior facilidade de recuperação de informação, a técnica de stemização foi empregada. O processo consiste em reduzir palavras flexionadas ao seu tronco principal, de forma que conjugações de verbos se aproximem mais do verbo no infinitivo, por exemplo (JIVANI, 2006, p.1930).

REPRESENTAÇÃO

O modelo de representação de texto utilizado foi o *Bag of Words.* O método, em seu caso mais simples, produz uma matriz esparsa de frequências dos termos contidos em todo o conjunto de dados (RADOVANOVIĆ; IVANOVIĆ, 2008, p. 228). Dessa maneira, cada documento é representado por um vetor que contém a

contagem de todos seus n-gramas, com uma frequência nula para os termos sem ocorrências.

Entretanto, a frequência absoluta da ocorrência de um termo pode ser tendenciosa. Termos recorrentes dentro do conjunto de documentos podem ser menos informativos quando comparados com termos esporádicos.

Para lidar com essa característica, foi utilizado o valor *tf-idf*, produto das duas medidas estatísticas *term frequency (tf)* e *inverse document frequency (idf).* O valor *tf-idf* reduz o impacto de *tokens* que ocorrem frequentemente em um dado corpus, pois são empiricamente menos informativos que aquele que ocorrem em uma menor porção do corpus de treinamento.

O valor *tf-idf* utilizado é obtido através da Eq. (1), onde $\mathrm{tf}(t, d)$ é a contagem absoluta das ocorrências do *token* t no documento d.

$$\text{tf-idf}(t, d) = tf(t, d) \times idf(t) \tag{1}$$

O valor de $idf(t)$ é obtido através da Eq. (2), onde n é o número de documentos que constituem o corpus, e $\mathrm{df}(t)$ é definido como o número de documentos da coleção que contém o termo t.

$$idf(t) = log\,\frac{n}{df(t)} \tag{2}$$

TREINAMENTO E VALIDAÇÃO

Com o objetivo de obter um modelo capaz de identificar o posicionamento expressado no discurso, analisando apenas sua transcrição, fez-se uso de métodos de aprendizado de máquina supervisionado para treinar classificadores.

Além da divisão treino/teste, o método de validação cruzada *k-fold* foi empregado com $k = 10$. Ao final do treinamento, o modelo que obteve os melhores resultados foi avaliado no conjunto de teste, com o objetivo de realizar uma verificação final da qualidade do modelo.

Para identificar potenciais anomalias no *dataset*, as instâncias classificadas incorretamente pelo modelo final passaram por inspeção manual com o objetivo de averiguar a coerência entre o conteúdo do texto e sua classe. O procedimento foi realizado durante a etapa de treinamento cruzado, bem como para a validação final com o conjunto de teste. Portanto, todas as instâncias do conjunto de treino também foram consideradas.

RESULTADOS E DISCUSSÃO

A métrica de avaliação utilizada para medir quantitativamente a qualidade dos modelos finais foi o *F1-Score*, e os resultados numéricos são apresentados na Tabela 1. A tarefa de classificação apresentou resultados promissores, com altos valores de F1, indicando a robustez do método. Em particular, o classificador bayesiano *Multinomial Naive Bayes* se aproximou de atingir o valor máximo da métrica utilizada, apresentando uma taxa de erro de apenas 1%.

Os resultados indicam que, apesar de as classes não serem facilmente separáveis através de inspeção visual, conforme discutido anteriormente, os classificadores foram capazes de identificar o posicionamento considerando apenas a representação do texto original.

Após inspeção manual dos dados que obtiveram classificação incorreta (11), apenas um texto apresentou indícios de comportamento anômalo ou ambíguo. Os outros 10 textos não apresentaram motivo aparente para o erro de classificação.

Tabela 1 – Comparação quantitativa entre classificadores utilizados

Classificador	F1-Score
Support Vector Machine	0.945
Multinomial Naive Bayes	0.990
Logistic Regression	0.923
K-Nearest Neighbors	0.786

Fonte: Autoria própria.

O desempenho inferior do classificador *K-Nearest Neighbors* pode ser explicado pelo fato de que, no espaço de *features* (Figura 1), os dados não aparentam formar dois grupos distintos, homogêneos e bem definidos. Portanto, espera-se que o método, que utiliza a distância dos k pontos mais próximos, apresente resultados menos promissores, principalmente devido à alta dimensionalidade dos dados.

CONCLUSÃO

Os resultados de classificação mostram que os grupos são facilmente separáveis com a representação adequada dos dados e aplicação dos métodos apropriados. Isso sugere que os discursos de uma mesma classe são, de maneira geral, agnósticos à partido e orador, sendo semelhantes entre si independente destes fatores.

O método também se mostra promissor para encontrar anomalias em bases de dados textuais, apesar de a quantidade de classificações incorretas ser pequena. Entretanto, devido ao tamanho do banco de dados utilizado, e pelo fato de não haver conhecimento prévio a seu respeito, os resultados não foram totalmente conclusivos. A utilização de um conjunto de dados maior e mais diversificado, do qual se tem conhecimento prévio sobre instâncias incorretamente rotuladas, poderia confirmar a hipótese.

Por fim, trabalhos futuros podem explorar representações de texto e métodos mais sofisticados, como word2vec e redes neurais, além de utilizar *datasets* mais diversificados e adequados. Com uma representação adequada, métodos de aprendizado não supervisionado, como de agrupamento, podem se provar opções viáveis.

AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa de Iniciação Científica (Edital PROPPG 02/2019 — PIBIC) concedida ao acadêmico Vinícius Couto Tasso.

REFERÊNCIAS

ROBERTO, José et al. Ajuste fiscal e reforma previdenciária. **Revista Conjuntura Econômica**, v. 73, n. 4, p. 23-25, 2019.

KUMAR, Vivek et al. Anatomy of preprocessing of big data for monolingual corpora paraphrase extraction: source language sentence selection. In: **Emerging Technologies in Data Mining and Information Security**. Springer, Singapore, 2019. p. 495-505.

RAULJI, Jaideepsinh K.; SAINI, Jatinderkumar R. Stop-word removal algorithm and its implementation for Sanskrit language. **International Journal of Computer Applications**, v. 150, n. 2, p. 15-17, 2016.

JIVANI, Anjali Ganesh et al. A comparative study of stemming algorithms. **Int. J. Comp. Tech. Appl**, v. 2, n. 6, p. 1930-1938, 2011.

RADOVANOVIĆ, Miloš; IVANOVIĆ, Mirjana. Text mining: Approaches and applications. **Novi Sad J. Math**, v. 38, n. 3, p. 227-234, 2008.