METODOLOGIA DE IDENTIFICAÇÃO DE POLARIDADE EM TEXTOS COM BASE EM PROJETOS DE LEI BRASILEIROS

Daniel L. B. Dos Santos [1]

Thais M. C. R. de Camargo [2]

Myrian C. A. Costa [1]

Valeria M. Bastos [3]

Nelson F. F. Ebecken [1]

daniellopes@coc.ufrj.br

thaismcrcamargo@usp.br

myrian@ntt.ufrj.br

valeriab@dcc.ufrj.br

nelson@ntt.ufrj.br

[1] PEC-COPPE/UFRJ

[1] Av. Athos da Silveira Ramos, 149 Centro de Tecnologia - Bloco B, Sala 101 - Ilha do Fundão, 21941-909, Rio de Janeiro, RJ, Brasil

[2] Programa de Pós-Graduação em Ciência Política/USP

[2] Av professor Luciano Gualberto, 315 – Cidade Universitária, 05508-900, São Paulo, SP, Brasil

[3] DCC/UFRJ

[3] Av. Athos da Silveira Ramos, 247 CCMN – Bloco E, sala E2016 – Ilha do Fundão, 21941-916, Rio de janeiro, RJ, Brasil

Abstract. *O objetivo deste trabalho é apresentar a pesquisa de uma metodologia eficiente de identificação de polaridade de documentos sobre um mesmo tema, usando como estudo de caso projetos de lei contra e a favor da liberalização do aborto no Brasil. Busca-se primeiramente identificar e comparar técnicas de aprendizado de máquina sobre textos que consigam classificar os projetos com vieses a favor e contra o aborto. Esta metodologia analisa cada parte do projeto em questão, utilizando a técnica de agrupamento K-means, aplicando*

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

em seguida um método baseado em grafos para processar todas as combinações de parâmetros e verificar os projetos que possuem maiores ligações entre si. Para a realização de testes foram utilizados projetos previamente classificados e os resultados obtidos com esta metodologia demonstraram ligações e peculiaridades muito interessantes e promissoras, ajudando assim, na identificação de similaridades e padrões em documentos.

Keywords: *Aprendizado de Máquina, Mineração de Textos, Clusterização, Polarização de Textos*

1 INTRODUÇÃO

A Política vem se tornando um dos tópicos mais discutidos e estudados no Brasil e no mundo. Consequentemente surge a necessidade de se obter informações sobre projetos e ações do executivo e do legislativo, bem como os responsáveis por estes. Esta necessidade se apresenta tanto para a informação geral da população quanto para estudos acadêmicos. Mas, devido ao grande número de documentos disponíveis, tais como os projetos de lei, sendo estes às vezes extensos, torna-se fundamental a utilização de ferramentas que automatizem as análises e sintetizem essas informações viabilizando a democratização do seu acesso. Desta maneira será possível separar os projetos por temas, e dentro de cada tema identificar suas polaridades, ou seja, se são contra ou a favor, aumento ou diminuição de conteúdo, entre outros. O foco deste trabalho é identificar polaridade em projetos de lei que possuem como tema o aborto. Define-se como polaridade a diferenciação dos projetos em relação ao seu posicionamento sobre o tema. Assim, busca-se separar os documentos em dois grupos, favoráveis e contrários a liberalização do aborto. Além disso, visa-se identificar as principais palavras e termos utilizados em cada grupo, bem como o quão interligados os projetos estão quanto ao seu tema e conteúdo.

Para este estudo foi utilizado o método de agrupamento K-means (Miner et al., 2012), implementado na ferramenta scikit-learn (K-means, 2017). Este método foi escolhido para que fosse possível analisar o comportamento dos projetos e suas ligações. Posteriormente, estes resultados serão comparados com outros modelos supervisionados e não supervisionados.

Este estudo visa desenvolver uma metodologia eficiente que possa ser utilizada para polarização de projetos de lei relacionados a qualquer tema, e não somente ao do estudo de caso levantado neste artigo. Um exemplo dessa utilização seria a de especialistas em ciência política, que não precisarão mais classificar manualmente projetos de lei quanto a sua polaridade e nem identificar as palavras encontradas que reforcem cada classificação. Através dos resultados encontrados neste estudo, os projetos contrários ao aborto podem servir de base para "tomada de decisão" que sejam favoráveis à população, implementando políticas de melhoria da saúde e diminuição de violência à mulher.

Este artigo está dividido em 5 seções, onde na seção 2 é levantado o estudo de caso, descrevendo a base de textos utilizada nesta análise e suas características. Na seção 3 é descrita a metodologia utilizada. A análise dos resultados é apresentada na seção 4 e a conclusão e trabalhos futuros na seção 5.

2 ESTUDO DE CASO

Para este estudo 33 projetos de lei brasileiros (PLs) foram fornecidos por uma especialista do Programa de Pós-Graduação de Ciência Política da USP, cujo estudo é focado no tema aborto e projetos de lei que o abordam. Juntamente aos projetos foi fornecida uma tabela, semelhante à mostrada na Tabela 1, representando a polaridade de cada documento, rotulados de forma manual pela própria especialista. Na tabela 2 são apresentadas as quantidades de PLs presentes em cada polaridade.

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

Inicialmente, os projetos de lei foram analisados manualmente, um a um, para se verificar sua estrutura. Os PLs seguem certo padrão com relação a sua apresentação, podendo assim ser dividido em três partes: uma breve descrição sobre o assunto que a lei tratará; o "Decreto" em si, que enuncia a lei a ser proposta ou uma alteração de lei já vigente; e, por último, uma "Justificativa", contendo os argumentos através de um texto corrido em favor da proposta apresentada. O início da "Justificativa" no PL é definido por um subtítulo.

Entretanto, eles não seguem fielmente estas definições de forma tão padronizada, dificultando um pouco o processamento textual. Por exemplo: Alguns projetos se referem à "Justificativa" como "Justificação" em seu subtítulo; outros omitem o subtítulo do tópico, diferenciando do texto do decreto apenas pela quebra de página; alguns outros modificam a forma como a palavra é representada, incluindo espaços entre cada letra ("J U S T I F I C A Ç Ã O").

Tabela 1. Amostra de projetos de lei rotulados

Projetos de lei	A favor ou contra o aborto
PL4834_2005	A favor
PL5364_2005	A favor
PL478_2007	Contra
PL489_2007	Contra
PL660_2007	A favor

Tabela 2. Quantidade e polaridade rotulada dos projetos de lei

Polaridade	Número de projetos de lei
A favor do aborto	13
Contra o aborto	20

Considerando os 33 exemplares adquiridos para este estudo, contornar este problema é relativamente simples. Visando trabalhos futuros, onde um montante bem maior de projetos poderá ser tratado, essa falta de padronização na estrutura do projeto pode se tornar uma dificuldade relevante.

Outra dificuldade identificada ainda no levantamento de dados a serem analisados foi com relação aos decretos. Alguns apenas modificam uma lei, removendo um inciso, ou até revogando uma lei inteira. Nestes casos, estes projetos acabam não possuindo informações relevantes na análise textual, dificultando seu processamento para identificação de polaridade.

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering
P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

Justificações muito pequenas ou sem relevância também representam um empecilho neste processamento. Quanto menos informação mais difícil se torna a tarefa de polarização, principalmente na justificação, onde o texto é menos imparcial e onde o autor pode argumentar e demonstrar suas ideias e convicções.

3 METODOLOGIA

Diversos estudos sobre estado da arte relacionados ao propósito deste artigo foram analisados, como o de Liu et al. (2015), que identifica palavras e termos relevantes para posteriormente serem rotulados e agrupados, e de Sista & Srinivasan (2004) que propõe um método de polarização léxica para classificação de textos. Porém, devido à especificidade deste tema, principalmente por tratar textos em português e estes textos serem projetos de lei brasileira, foi definido não se utilizar dos métodos anteriores e se testar e analisar uma nova abordagem, utilizando-se de ferramentas já reconhecidas de classificação de textos e identificação de termos relevantes e similares. As sessões a seguir descrevem o processo de pré-processamento dos projetos de lei, seu agrupamento através do algoritmo K-means e formas de representação para a visualização e melhor análise destes resultados.

3.1 Pré-Processamento

Os PLs foram processados automaticamente e salvos em memória separando o "decreto" e a "justificativa". Para isto, foram utilizados os padrões identificados manualmente e suas exceções, como exemplificado na seção 2.

Neste estudo, o "decreto" e a "justificativa" foram processados, tanto juntos como de maneira individual, para que se avalie o impacto de cada um e o quão cada parte é mais significativa para a polarização. Desta forma, essas três possibilidades de texto são utilizadas como descrito na tabela 3, dependendo do valor da variável chamada de "Textual". Por exemplo, se "Textual" for 1, somente o decreto é processado e analisado.

Tabela 3. Valores da variável "Textual"

Valor da variável "Textual"	**O que representa**
0	O Projeto todo – decreto e justificativa
1	Somente o decreto
2	Somente a justificativa

Todos os PLs foram processados utilizando a linguagem Python (Python, 2017) e suas bibliotecas, onde todas as palavras foram reduzidas a forma caixa baixa (letras minúsculas), e salvos de forma estruturada, separando a sua justificativa do decreto. Os projetos foram lidos e armazenados em memória de forma ordenada em relação ao ano e nome. Expressões regulares

foram utilizadas para tratar os textos e estruturá-los para armazenamento, além de contornar as dificuldades sobre a falta de padronização, descrito anteriormente.

Em seguida foram removidas as *stopwords*, utilizando uma lista existente na biblioteca do NLTK (Bird et al., 2009). Os PLs foram representados através de *bag-of-words*, com termos no formato N-gram, onde N varia de 1 a 4 palavras, gerando assim, uma matriz PL x Termo. Para o cálculo do peso dos termos foi utilizado o método TF-IDF (Miner et al., 2012) nativo do scikit-learn (2017), representado pela Eq. (1).

$$TF \times IDF = tf(t,d) \times idf(t) = tf(t,d) \times \left(log \frac{1+n_d}{1+df(d,t)} + 1 \right) \tag{1}$$

onde $tf(t,d)$, *term frequency*, representa o número de vezes que o termo t aparece no documento d, n_d representa o número total de documentos e $df(d,t)$ representa o numero de documentos que possuem o termo t. Foi ativado um parâmetro chamado "smooth_idf" do método do scikit-learn, que é responsável por somar 1 ao numerador e denominador, a fim de se prevenir a divisão por zero.

Num segundo passo foi realizada a normalização de cada linha (vetor *v*) utilizando a distância euclidiana (Eq. (2)), nativa do scikit-learn. Desta maneira, foi obtida uma matriz de PLs por termos.

$$v_{norm} = \frac{v}{\|v\|_2} = \frac{v}{\sqrt{v_1^2+v_2^2+\cdots+v_n^2}} \tag{2}$$

Neste método, foi definida como parâmetro uma variável "N-gram" com "N" único, não tendo assim no *bag-of-words* N-grams de Ns diferentes. "N" variando de 1 a 4.

3.2 Agrupamento

O K-means é um método de agrupamento no qual os elementos são reagrupados a cada iteração, recalculando o centroide até atingir o equilíbrio dos elementos em cada grupo. Não foi considerado nenhum valor inicial para a semente do algoritmo. Ao invés disso, foi utilizado o método *K-means++*, sugerido por Arthur & Vassilvitskii (2007) e implementado no scikit-learn, que modifica a forma de inicialização dos centroides. No algoritmo padrão os K centroides iniciais são escolhidos de forma randômica ou são definidos na entrada do algoritmo pelo usuário. Desta forma o algoritmo converge, dado um tempo de execução suficiente, mas eles podem convergir para mínimos locais, devido a proximidade das sementes. O método *K-means++* tenta minimizar esse problema, definindo centroides geralmente mais distantes entre si. Este algoritmo é definido abaixo:

1. Escolha dos centroides iniciais:
    a. Define-se um centroide c_1, escolhido uniformemente de forma aleatória de X.

b. Define-se um novo centroide c_i, escolhendo $x \in X$ com probabilidade $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$.

c. Repete-se o passo 1.b até se ter todos os k centroides.

2. Passos normais do K-means:

a. Para cada $i \in \{1, \ldots, k\}$, o cluster C_i é definido como o conjunto de elementos (pontos) de X que são mais próximos de c_i do que c_j, para todo $j \neq i$.

b. Para cada $i \in \{1, \ldots, k\}$, define-se como c_i o centro de massa de todos os pontos em C_i: $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.

c. Repete-se os passos 2.a e 2.b até C não se modificar mais.

Além das variáveis "Textual" e "N-gram", foram testados diferentes valores para o número de grupos do K-means, variando-se o "K" de 2 a 7. Nesta análise, foram combinadas as três variáveis citadas anteriormente: "Textual", "N-gram" e "K", gerando assim um total de 72 combinações de execução. Uma primeira avaliação foi a análise de quantos projetos negativos e positivos cada grupo de cada execução apresentava, onde, por exemplo, os favoráveis e contrários se isolassem em seus grupos. Na Fig. 1 é exemplificada a combinação {K=5; Textual=0; N-gram=4}, chamada de combinação C1.

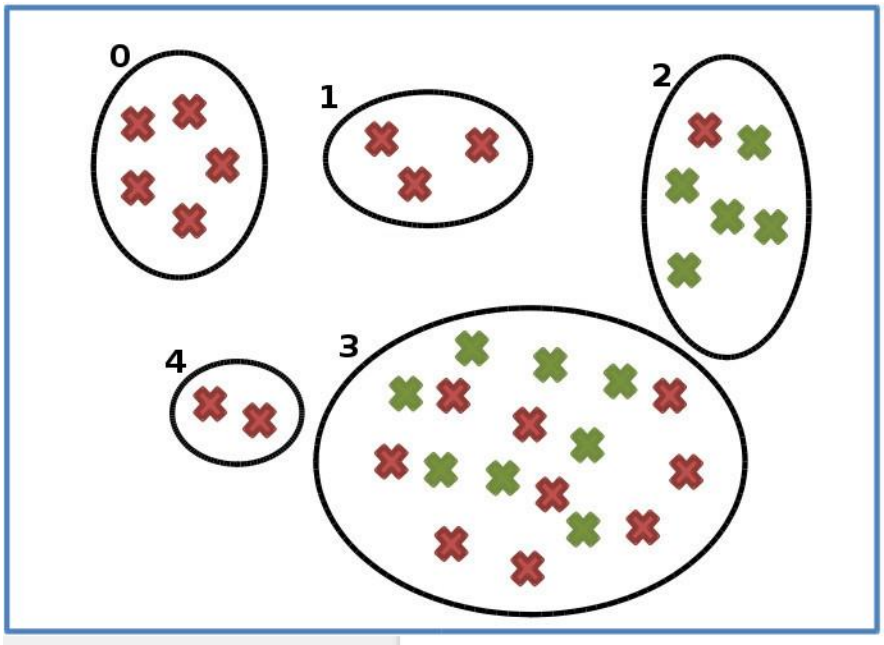

Figura 1. Análise de cenário da combinação C1{K=5; Textual=0; N-gram=4}

Os elementos vermelhos representam os projetos contra o aborto enquanto os verdes a favor. Nela é possível observar que o método separa perfeitamente os projetos nos grupos 0, 1 e 4. No grupo 2 é verificado somente um PL discrepante do resto de seus elementos. O grupo 3, entretanto, que contém a maior quantidade de projetos, sendo nove contra e oito a favor,

possui uma certa similaridade entre seus termos. Foi verificado que na maioria das combinações de parâmetros, no mínimo um agrupamento apresentava este mesmo comportamento.

Um aspecto interessante a se analisar é a ocorrência de PLs juntos em um mesmo grupo, independente de sua polaridade pré-definida, no decorrer de várias execuções. Estes projetos podem apresentar características semelhantes que os coloquem juntos em diversas execuções do algoritmo com parâmetros distintos.

O alto número de combinações dificulta a análise individual dos gráficos ou dados em busca das similaridades entre os PLs. Foram desenvolvidas outras formas de representação destes resultados.

3.3 Representação em grafos

Com objetivo de visualizar melhor estas ligações recorrentes entre PLs, foi proposta a representação em grafos, utilizando a biblioteca igraph em Python (igraph, 2017). Na fig. 2 é apresentado um exemplo de grafo, correspondente à combinação C2 {K=(5:6), Textual=0; N-gram=4}, onde "K=(5:6)" significa que o grafo representa as execuções com "K"=5 e "K"=6.

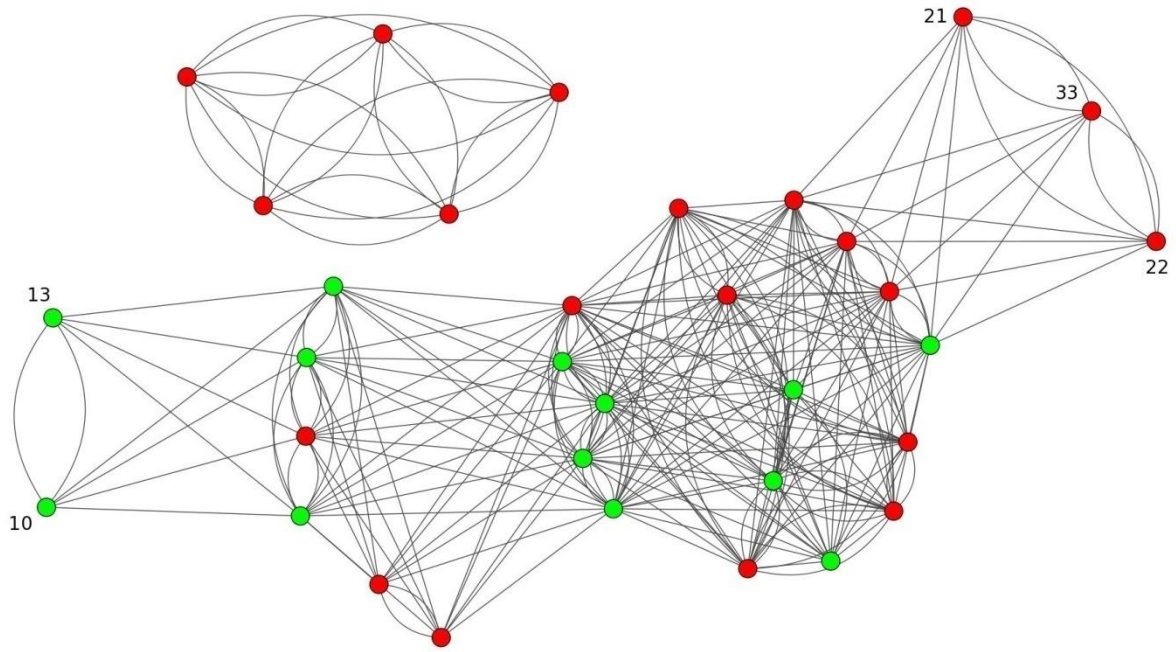

Figura 2. Representação em grafo da combinação C2 {K=(5:6), Textual=0; N-gram=4}

Os vértices dos grafos representam os PLs e as arestas significam que dois PLs se encontram em um mesmo grupo em, pelo menos, uma combinação de parâmetros. Por exemplo, no grupo 1 da combinação C1, demonstrada na fig. 1, cada PL possui 2 arestas, ligando-se aos outros 2 elementos do grupo. A cor dos vértices representa o rótulo de polaridade de cada projeto, fornecido pela especialista. Projetos contra o aborto receberam a cor vermelha e a favor receberam a cor verde. Esta solução trouxe um avanço na análise em alguns aspectos, tal como, a similaridade de comportamento e destaque de alguns PLs. Os PLs 13 e 10 apresentam um comportamento semelhante, ligando-se entre si e aos mesmos vértices. O mesmo vale para os

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

PLs 21, 22, 33. Estas ligações foram relatadas à especialista, que observou que os PLs (10 e 13) e (21 e 33) tratam-se de projetos equivalentes entre si. “Projetos equivalentes” são projetos que foram reescritos, ou submetidos novamente para apreciação, utilizando a mesma proposição de lei e na maioria das vezes a mesma justificativa ou uma similar, reforçando a anterior. Ou seja, estes projetos possuem praticamente o mesmo conteúdo, mudando somente o seu identificador e as palavras utilizadas.

Na fig. 3 é destacado o subgrafo isolado do grafo da fig. 2 com os respectivos nomes e índices dos PLs. Estes 5 PLs tiveram apenas ligações entre si. Outra observação relevante neste subgrafo é que todos possuem rótulos contra o aborto. A partir de grupos como este é possível analisar detalhes comuns e assim utilizá-los para identificar grupos de mesma polaridade. Uma análise mais detalhada dessas ligações é discutida na seção 3.

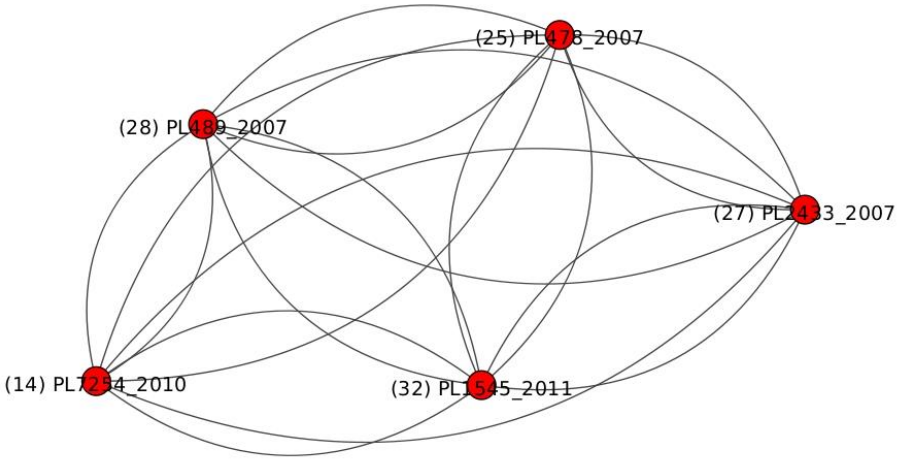

Figura 3. Destaque do grupo independente da fig. 2

O problema desta forma de visualização é o alto número de arestas, que dificulta a verificação de algo relevante de forma manual. Para uma avaliação mais abrangente de todas as ligações entre PLs computadas, foi gerado um grafo contendo as 72 combinações de execução. Um exemplo da dificuldade citada é mostrado na fig. 4 que representam estas 72 combinações de execução, ou seja, todas as execuções, com “K” variando de 2 a 7, “N-gram” de 1 a 4 e “Textual” de 0 a 2, chamada de combinação C3.

Uma forma de diminuir estas arestas, facilitando a visualização do grafo, foi unificar arestas iguais, ou seja, que ligam um vértice A ao B, e apresentá-las em apenas uma aresta ponderada representando o número de ligações presentes entre estes dois vértices. A fim de se destacar vértices que possuem mais ligações entre si, a ponderação é representada pela espessura da aresta, ou seja, quanto maior o número de ligações, maior a espessura.

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

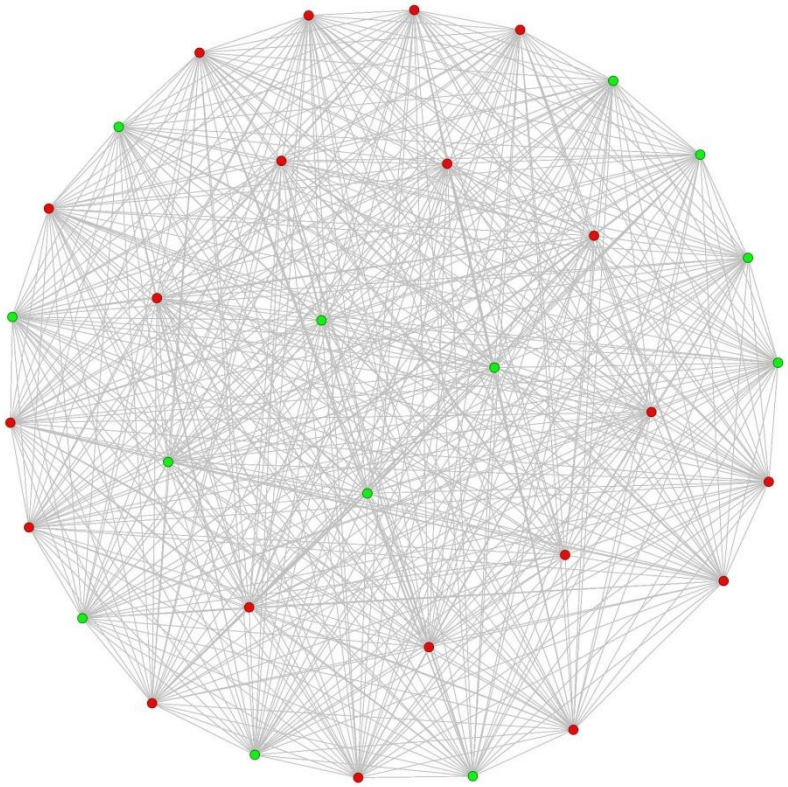

Figura 4. Grafo representando a combinação C3 {K=(2:7); Textual=(0:2); N-gram=(1:4)}

Foram gerados também outros grafos utilizando um percentual de corte destas ponderações, chamados "top grafos". Nestes grafos as ponderações abaixo deste corte são eliminadas. Por exemplo, considerando a combinação C3 {K=(2-7);Textual=(0-2);N-gram=(1-4)} com um corte de 100%, o grafo somente apresenta os vértices que possuem as ligações com o valor máximo (72).

Na combinação C3 com o corte de 80% (57 ligações ou mais), apresentada na Fig. 5a, pode ser observado que todas as ligações de PLs possuem mesma polaridade. Por outro lado, para a combinação C3 com corte de 70% (50), Fig. 5b, algumas ligações entre PLs com polaridades opostas ocorrem, junto com casos de ligações entre PLs semelhantes. Um recorte com aproximação para melhor entendimento está mostrado na Fig. 5c. As análises destes comportamentos estão relatadas na próxima seção.

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering
P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

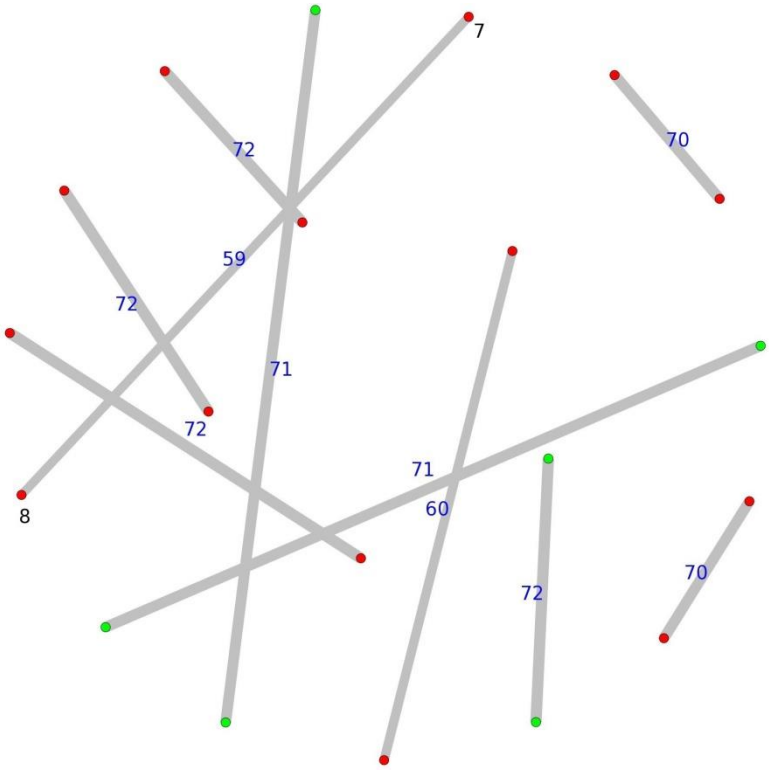

Figura 5a. Combinação C3 {K=(2-7);Textual=(0-2);N-gram=(1-4)} com o corte de 80%

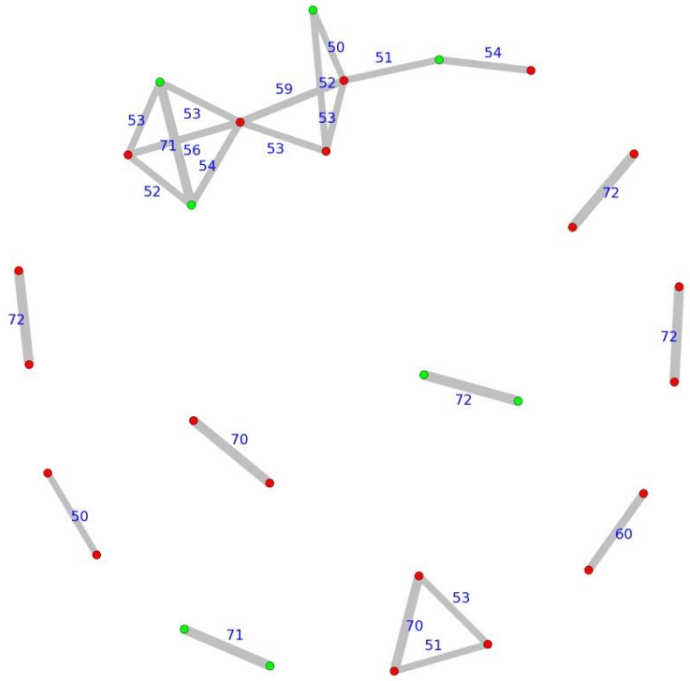

Figura 5b. Combinação C3 {K=(2-7);Textual=(0-2);N-gram=(1-4)} com o corte de 70%

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

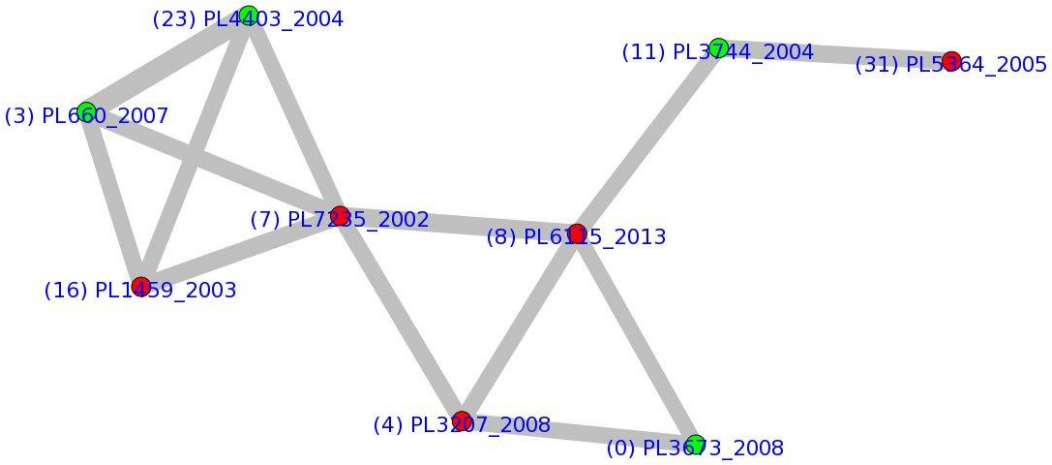

Figura 5c. Destaque do grupo independente da Fig. 5b

4 ANÁLISE DOS RESULTADOS

A análise em grafos dos resultados do K-means através das maiores ligações proporcionou levantamentos bastante relevantes sobre os projetos de lei. Os resultados iniciais e alguns grafos representando as maiores ligações foram analisados pela especialista para que fosse identificado, através do conhecimento prévio sobre os projetos, se existia algum motivo claro para o alto nível de ligações entre eles. Foi constatado que os projetos presentes no corte 100% da combinação C3 são projetos equivalentes.

Na fig. 2, como inicialmente relatado na seção 2.3, são apresentados os projetos equivalentes, onde o primeiro par (10 e 13) aborda como assunto comum a criação de uma bolsa auxílio para mulheres que decidirem realizar aborto após estupro ou que sofreram aborto espontâneo. O segundo par (21 e 33), por sua vez, discorre sobre incluir um dispositivo na Lei nº 10.406, de 10 de janeiro de 2002 - Código Civil (Código Civil, 2017), dispondo sobre o registro público da gravidez. Estes pares, inclusive, são exemplos de PLs equivalentes totalmente idênticos, possuindo exatamente o mesmo texto. Ambos os pares estão presentes também na combinação C3 com corte 100%. O PL 22, apesar de não ser um projeto equivalente aos outros (21 e 33), trata também a questão da obrigatoriedade do cadastramento da gestante, demonstrando assim que a análise também conseguiu identificar projetos que possuíam mesmo tema e polaridade.

Na fig. 3 pode-se destacar que todos os PLs possuem viés negativo, sendo que os dois mais acima ([28] PL478_2007 e [25] PL489_2007) correspondem ao projeto do estatuto do nascituro, que contém artigos similares do art. 2º do Código Civil. Ambos também estão representados como um par na combinação C3 com corte 100% e, consequentemente, podem ser vistos na fig. 5b. Os outros 3 projetos (14, 27, 32) possuem o mesmo objetivo, que é aumentar a pena de reclusão estipulada no Código Penal (Código Penal, 2017) a terceiros que realizaram (ou provocaram) o aborto. Os PLs 14 e 27 são projetos equivalentes, apresentando o mesmo texto, com exceção de dois artigos adicionais no decreto (124 e 124-A), referentes ao aumento da pena de reclusão também às gestantes que provocaram ou consentiram o aborto e a tipificação do crime de induzir ou auxiliar a gestante à prática do aborto, respectivamente. Já o PL 32 concentra-se no aumento da pena para médicos, incluindo a proibição definitiva dos

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering
P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

mesmos exercerem a profissão. A forte ligação entre estes três PLs é vista na Fig. 5b, onde eles formam um triângulo isolado.

Os projetos apresentados na fig. 5a são projetos equivalentes, com exceção dos projetos 7 e 8, que tratam do aborto em casos de estupro, mas são propostas distintas: O primeiro suprime a possibilidade de aborto nesses casos, enquanto o segundo busca estabelecer uma barreira ao acesso ao aborto no SUS (com a obrigatoriedade do corpo de delito). Este é mais um exemplo onde o estudo conseguiu identificar projetos de mesmo tema (e mesma polaridade) ou objetivo, apesar de possuírem características e maneiras diferentes de abordagem.

A análise do grafo da fig. 5c é mais complexa, pois os PLs apresentam fortes ligações entre si, apesar de possuírem polaridades distintas. Por exemplo, os projetos 0 (PL3673_2008) e 4 (PL 3207_2008) alteram a natureza do crime de aborto, mas com polaridades opostas: o primeiro o transforma em crime de "menor potencial ofensivo" e o outro o transforma em crime hediondo, além do induzimento, instigação ou auxílio ao suicídio. Já o projeto 8 (PL 6115_2013) cria barreiras ao aborto em caso de estupro (com a obrigatoriedade do corpo de delito). Outra análise interessante pode ser feita a partir dos PLs: 31 (PL5364_2005) e 11 (PL 3744_2004), pois os projetos possuem polaridades distintas. Ambos tratam da possibilidade de aborto em caso de estupro, porém, um expande essa possibilidade (incluindo atentado violento ao pudor) e o outro elimina essa possibilidade. Os projetos 3 (PL660_2007) e 23 (PL 4403_2004) são o mesmo projeto, que propõem autorizar o aborto em casos de anencefalia. O projeto 16 (PL1459_2003) torna crime o aborto em caso de anomalia fetal, ou seja, mesmo tema do par citado anteriormente, mas com diferente polaridade.

5 CONCLUSÃO E TRABALHOS FUTUROS

O estudo apresentado neste trabalho busca o desenvolvimento de uma metodologia eficiente de identificação de polaridade de documentos de projetos de lei. Foi utilizado um algoritmo de agrupamento que se mostrou bastante adequado ao propósito, e foi desenvolvida uma ferramenta eficiente de visualização dos resultados em grafos.

Foram encontrados projetos semelhantes e projetos com polaridades distintas, porém tratando do mesmo assunto, o que torna esse ambiente propício para a identificação de padrões de similaridade entre PLs.

Visa-se, posteriormente, utilizar toda a análise feita com os grafos e os PLs que possuem maiores ligações para identificar a polarização dos mesmos. Uma técnica levantada é a de se obter em cada grupo de polaridade, pré-classificado pela especialista, os principais N-grams, dando maior valor para os que mais aparecem nos grupos dos "top grafos" em destaque. Esta ação ajudaria na desambiguação de grupos como o 3 da Fig. 1. Alguns estudos sobre esta técnica já foram feitos, como em Amorim (2011).

Outros trabalhos futuros são a utilização de métodos supervisionados, em especial o SVM, nativo do scikit-learn (2017), aplicando também a visualização de sua classificação em grafos, e o teste de ponderação de N-gram mais relevantes de polarização, utilizando como base estudos anteriores como o de Joachims (1998) e Sista & Srinivasan (2004). Com isso, será possível comparar os resultados de ambos os métodos e analisar suas peculiaridades.

REFERÊNCIAS

Amorim, R., 2011. *Learning Feature Weights for K-means Clustering Using the Minkowski Metric.* PhD thesis, Birkbeck, University of London.

Arthur, D. & Vassilvitskii, S., 2007. K-means++: The advantages of careful seeding. In Gabow, H., eds, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA '07)*, pp. 1027-1035.

Câmara dos deputados. 2017. Disponível em: <http://www2.camara.leg.br/agencia/noticias/69896.html>. Acessado em: 03/07/2017

Código Civil, Lei nº 10.406. 2017. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/2002/L10406.htm>. Acessado em: 03/07/2017

Código Penal, Decreto-Lei nº 2.848. 2017. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto-lei/Del2848.htm>. Acessado em: 03/07/2017

Bird, S., Klein, E. & Loper, E., 2009. *Natural Language Processing with Python*. Disponível em: <http://www.nltk.org/book>. Acessado em 02/03/2017.

igraph, versão 0.7.0 para Python. Disponível em: <http://igraph.org/python/>. Acessado em: 02/08/2017

Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning With Many Relevant Features. In Nédellec, C. & Rouveirol, C., eds, *Machine Learning (ECML-98)*, pp. 137-142.dd

Liu, J., Shang, J., Wang, C., Ren, X. & Han, J., 2015. Mining Quality Phrases from Massive Text Corpora. In Sellis, T., Davidson, S. B. & Ives, Z., eds, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15), pp 1729-1744.*

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R. & Delen, D., 2012. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Elsevier.

Python, 2017. Disponível em: <https://www.python.org/>. Acessado em: 05/08/2017

scikit-learn. 2017. Disponível em: <http://scikit-learn.org>. Acessado em: 05/08/2017

Sista, S. P. & Srinivasan, S. H., 2004. Polarized Lexicon for Review Classification. In Arabnia, H. R. & Mun, Y., eds, *Proceedings of the International Conference on Artificial Intelligence (IC-AI '04)*, pp. 867-872. CSREA.

SVM, Support Vector Machines, 2017. Disponível em: <http://scikit-learn.org/stable/modules/svm.html>. Acessado em: 07/06/2017