RELATÓRIO TÉCNICO II

ANÁLISE RECLAMAÇÕES

28 DE FEVEREIRO DE 2020 // VERSÃO PRELIMINAR

PROJETO PESQUISA: ESTUDO, DESENVOLVIMENTO E APLICAÇÃO DE METODOLOGIAS DE GESTÃO DE INTEGRAÇÃO DE DADOS PARA ANÁLISE DE RELAÇÕES DE CONSUMO NO SETOR DE TELECOMUNICAÇÕES NO BRASIL

ANATEL ibict MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES PÁTRIA AMADA BRASIL GOVERNO FEDERAL

RELATÓRIO TÉCNICO PRELIMINAR - PROCESSAMENTO DE TEXTOS DE RECLAMAÇÃO

1. Resumo

Este relatório técnico preliminar documenta os principais resultados e conclusões desenvolvidos até a data de sua entrega. Este trabalho de processamento de linguagem natural aplicou diferentes técnicas de mineração de texto em textos de reclamação registrados na Anatel, a fim de se extrair novas percepções além das que hoje existem na árvore de classificação da Agência. Almeja-se que os resultados desenvolvidos até a versão final deste documento possam ser usados pela Anatel diretamente, com o compartilhamento do conhecimento desenvolvido, ou indiretamente, através do uso desses resultados, por exemplo, em análises preditas subsequentes.

2. Introdução

A análise LDA utilizada neste trabalho objetiva criar novas classificações textuais dos registros de reclamação, utilizando somente a informação presente no texto, para possibilitar uma extração de informação que possibilite novas visões sobre os dados além daquelas que as categorias atuais da árvore de classificação existente proporcionam. Assim, diferentes cenários podem ser explorados a partir deste trabalho, a depender dos (1) resultados e (2) objetivos.

Na técnica desenvolvida na análise, aplicou-se uma técnica de multi-classificação dos textos de reclamação, na qual, além da geração de novos tópicos (não supervisionados) foi possível reclassificar os textos como contendo um ou mais tópicos, em diferentes proporções. Assim, tópicos intermediários passam a poder ser interpretados junto a tópicos principais, permitindo a priori a identificação desses novos fatores.

O processo de interpretação dos tópicos é parte crucial em trabalhos de classificação não supervisionada, e grande cuidado deve ser tomado para que *insights* teóricos e práticos possam ser levados em consideração no momento da correta nomeação.

O relatório técnico preliminar está dividido da seguinte forma: na (3) *metodologia* é apresentado um breve resumo dos dados utilizados para análise do modelo; uma breve pontuação das principais técnicas utilizadas no processo de pré-processamento; a base da teoria LDA utilizada como referência; as técnicas de visualização utilizadas; e na (4) *interpretação da classificação textual de reclamações* é iniciada a discussão acerca da interpretação dos tópicos, e próximos caminhos são pontuados para desenvolvimento da análise.

3. Metodologia

A metodologia, embora não detalhada extensivamente, detalha os principais pontos para entendimento da metodologia LDA aplicada na pesquisa. Discussão detalhada referente a parâmetros utilizados, análises de sensibilidade e teste de coerência dos tópicos serão pontuados na versão final do documento. A seguir são apresentados os (1) Dados utilizados; (2) Pré-processamento; (3) Análise Latent Dirichlet Allocation ; (4) O conceito de tópicos latentes; (5) o modelo LDA ; (6) Aplicação LDA e (7) Visualização de resultados.

3.1. Dados utilizados

Os dados utilizados para a análise de reclamações são os dados da antiga base de dados da Anatel Focus (Atual Anatel Consumidor [1]) e compreendem dados de 2018. As únicas informações utilizadas no treinamento do modelo são os dados de reclamação em texto.

A fim de reduzir o viés do operador intermediário na origem dos dados, as reclamações registradas via *Call Center*; *Atendimento Pessoal* e *Outros* são retirados da análise no momento de treinamento, restando somente aqueles em que o consumidor é o responsável direto pelo registro. Reclamações da categoria *Longa Distância/Interurbano* e *Programa Banda Larga nas Escolas (PBLE)* também foram retirados por serem considerados pouco relevantes para a formação de um fator (em especial considerando o volume em declínio), não representando prejuízo para eventual aplicação posterior nestes casos.

3.2. Pré-processamento

O primeiro passo para preparar os dados para análise são: (1) Transformação de todos os caracteres para minúsculos; (2) filtro por apenas caracteres alfabéticos; (3) retirada de acentuações; (4) Tokenização[2]; (5) Lemmatization ;(6) aplicação de bigramas e trigramas; (7) filtro por termos com tamanho mínimo de dois caracteres e (8) Filtro de Stopwords[3].

Os passos (1), (2) e (3) são processados de maneira direta, e embora sejam processados individualmente, muitas bibliotecas de código aberto já permitem a sua aplicação automática. Os passos (4) e (5) são feitos com o uso da biblioteca SpaCy [4](EXPLOSION, 2017) e exploram o conceito de part-of-speech, no qual são filtradas apenas palavras com funções desejadas na frase (como verbos e substantivos) além de as normalizar para versões comuns (e.g. enviou e enviado viram enviar).

---

[1] Disponível em https://www.anatel.gov.br/consumidor/component/content/article/44-noticias/959-anatel-lanca-nova-plataforma-de-atendimento-anatel-consumidor .

[2] Transformar frases em “tokens”, ou simplesmente palavras individuais.

[3] Stopwords, ou Palavras Vazias. São palavras retiradas do texto, as quais podem ser interpretadas como relevantes para o modelo, quando são conhecidamente indesejadas/nulas (e.g. conectores textuais).

[4] Documentação disponível em https://spacy.io. Licença MIT.

No passo (6), os bigramas e trigramas são formados são conjuntos de palavras (dois ou três, respectivamente) que trazem juntas um sentido diferente que quando separadas, independente do contexto (e.g. água e viva, ou água-viva). O passo (8) são palavras vazias que não trazem contexto ou podem enviesar o modelo matematicamente por aparecem com alta frequência mesmo sem agregar significado. Essa passo é feito iterativamente com a ajuda de vocabulário prévio, vocabulário identificado empiricamente, ou conhecimento empírico dos termos sem valor semântico, porém frequentes.

3.3. Análise Latent Dirichlet Allocation

A principal referência técnica usada neste trabalho é a publicação de mesmo nome da técnica desenvolvida - Latent Dirichlet Allocation – (BLEI; NG; JORDAN, 2003). A técnica é uma das mais utilizadas em mineração de texto para *information retrieval*, e apesar de ter aplicações em outras áreas de conhecimento além do Processamento de linguagem natural, sua maior aplicação prática é na multi-classificação textual através de tópicos.

Diferentemente da técnica de tf-idf[5], na qual são contabilizadas as ocorrências de uma palavra em uma frase, divididos pela ocorrência total em outros documentos, a análise LDA tenta trazer mais contexto para a interpretação da semântica dos textos com tópicos latentes (não mensuráveis diretamente). O algoritmo não busca somente palavras, mas (principalmente) contextos similares, o que diminui a dependência por palavras específicas (já que não usa a frequência como principal referência, mas um plano linear no espaço vetorial, que envolve múltiplas palavras de um mesmo contexto) e permite a aplicação em textos novos (não usados no treinamento do modelo).

A técnica LDA é desenvolvida, principalmente, sobre dois conceitos até então já difundidos. O primeiro aplica técnica de redução de dimensionalidade no espaço vetorial , por exemplo, tf-idf de forma a se obter um espaço dimensional linear capaz de identificar palavras sinônimas: a análise LSI (Latent semantic indexing). O segundo conceito é o LSI probabilístico (ou pLSI), o qual considera as palavras como amostras de modelos mistos, nos quais os componentes dos modelos são interpretados como variáveis latentes/tópicos.

A análise pLSI, no entanto, apenas analisa as relações a nível de palavras (embora condicionalmente para cada tópico e documento), e o modelo é dependente do espaço amostral utilizado no treinamento, tendo aplicação em novos textos inserta. A análise LDA usa do conceito de intercambialidade a nível das palavras (a ordem das palavras não altera a análise) e a nível do documento, simultaneamente. O conceito de intercambialidade, no entanto, não se refere à independência (não significa que a ordem das palavras não traga mais informação), mas sim a independência condicional, referente aos parâmetros da distribuição de probabilidade (os quais permitem que a estrutura intra-documento seja inferida), os quais serão explicados ao longo da *metodologia*.

3.3.1. O conceito de tópicos latentes

[5] Term-frequency Inverse Document-Frequency.

Antes de explorar o conceito de tópicos latentes via LDA, é interessante analisar como se formam os tópicos estudados através de modelos mais simples. A seguir serão apresentados 3 modelos que ajudam a explicar a formação do modelo LDA: Unigramas; LSI e pLSI.

3.3.1.1. Modelo Unigramas

Mesmo que bigramas e trigramas tenham sido utilizados neste trabalho, o modelo de unigramas elucida o conceito inicial base dos demais modelos. No modelo de Unigramas, palavras (w) são unigramas independentes, os quais fazem parte de um documento/reclamação (N) em um vocabulário (M) (Figura 1).

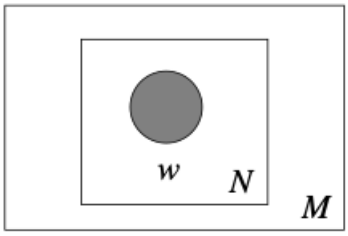


Figura 1 - Modelo Unigrama

Assim, a probabilidade de que uma palavra específica apareça (em qualquer contexto) é simplesmente dada pelo produtório da probabilidade desta palavra aparecer em cada um dos documentos usados no treinamento, seguindo uma mesma distribuição multinominal para todos os casos.

Equação 1 - Modelo unigrama

$$p(\mathrm{w}) = \prod_{n=1}^{N} p(w_n)$$

Essa simplificação resulta em um modelo que privilegia palavras com alta frequência, ou mesmo frequência elevado no *TF-IDF*. Contudo, o modelo de unigramas não nos dá informações de tópicos que possam ajudar a agregar essas palavras.

3.3.1.2. Modelo LSI

O modelo LSI trás o conceito de uma camada extra em relação ao modelo de unigramas, no qual são usados tópicos (z) para modelar mais de uma distribuição multinomial (Figura 2).

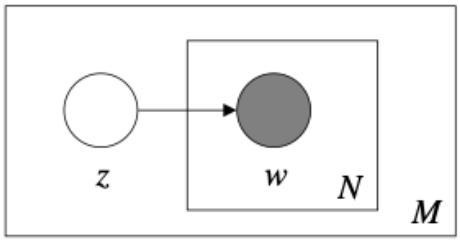


Figura 2 - Modelo LSI

Na prática, isso equivale a ter múltiplos modelos unigramas, dos quais para cada tópico z tem-se uma distribuição de palavras prováveis.

Equação 2 - Modelo LSI

$$p(\mathrm{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z)$$

Contudo, devido ao fato de as variáveis não dependerem do contexto em que estão (ou seja, estamos analisando apenas palavras, e tópicos a que elas possam participar), podemos tão somente chegar à conclusão do tópico mais provável para cada palavra, o que classificaria o texto em apenas um tópico (o que seria equivalente a identificar a o tópico em que a maior parte das palavras é classificada).

3.3.1.3. Modelo pLSI

O modelo probabilístico LSI trás uma camada de informação referente ao documento analisado individual. Isso significa que, ao treinar o modelo, damos a informação de quais palavras participam de qual documento, o que permite uma classificação também a nível de documento (Figura 3).

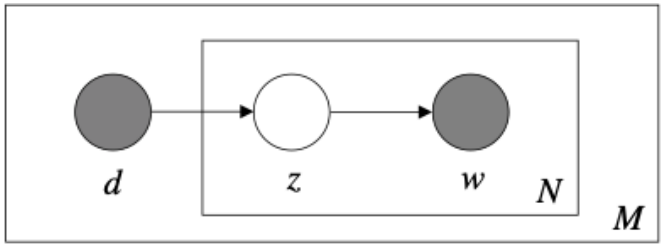


Figura 3 - Modelo pLSI

Assim, é utilizada a informação da probabilidade de aparecimento de uma palavra em todos os tópicos estudados, dado a probabilidade de ocorrência daquele tópico em um documento específico.

Equação 3 - Modelo pLSI

$$p(\mathrm{d},w_n) = p(d) \sum_{z} p(w_n|z)\, p(z|d).$$

Embora esse resultado seja o mais próximo do modelo LDA, permitindo uma classificação para cada documento do percentual a que cada tópico participa, o modelo pLSI tem duas grandes barreiras. Como todas as probabilidades são condicionais a nível de palavras e documentos, o custo computacional é elevado ou inviável a depender do volume de texto estudado. E segundo, como as informações de probabilidade a nível de documento são calculadas individualmente, o modelo só é capaz de ser aplicado nos mesmos textos em que foi treinado, não sendo possível classificar textos novos.

3.3.2. Modelo LDA

Por fim, o modelo LDA generaliza o modelo pLSI, ao trazer parâmetros da distribuição Dirichlet a priori sobre as já existentes distribuições (uma distribuição de distribuições). Isso permite que, ao invés de determinar os resultados baseado na otimização individual de cada reclamação, os valores probabilísticos sejam estimados através desta distribuição, inferindo o processo de formação das reclamações. As informações a priori da distribuição são passados a nível de tópicos por reclamação e palavras por tópicos (Figura 4) com os parâmetros Alpha e Beta (Teta é a distribuição Dirichlet de Alpha).

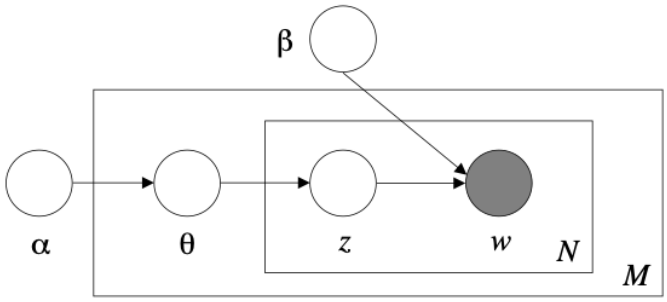


Figura 4 - Modelo LDA

O parâmetro Alpha controla a diferença entre tópicos (quanto maior, mais parecidos serão os tópicos uns dos outros), enquanto que o Beta controla a similaridade a nível de palavras dos tópicos (quanto maior, mais as palavras que formam os tópicos serão comuns). Contudo, ambos os parâmetros são estimados automaticamente em um processo de otimização, necessitando da definição, principalmente do número de tópicos.

Dessa forma, é possível inferir os resultados a nível de cada documento com uma redução computacional significativa e, principalmente, torna-se possível analisar textos novos. A definição da formulação, então, passa a ser:

$$p(\mathrm{w}|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta)\right) d\theta_d$$

Como temos agora a probabilidade marginal condicional aos parâmetros da distribuição Dirichlet, a integral faz com que possamos ter os valores discretos para cada caso.

3.4. Aplicação LDA

A principal biblioteca utilizada para a implementação da análise LDA foi a biblioteca python Gensim[6] (REHUREK; SOJKA, 2011). O modelo utilizado representa uma pequena variação em relação ao modelo original, denominado *Online LDA*, o qual usa de um processo de otimização estocástica para reduzir o custo computacional e convergir mais rapidamente os resultados, em especial quando o volume de texto é muito grande (HOFFMAN; BACH; BLEI, 2010).

De maneira geral, para uma análise lda, cria-se um dicionário de palavras numérico e, em seguida, um corpo de textos comum, usando o dicionário. Para a criação do corpo comum, utiliza-se da técnica de bolsa de palavras, na qual as palavras são traduzidas para números em colunas que contam suas frequências. Vale notar que esse processo faz com que as palavras fiquem em ordem diferente daquela em que foram produzidas.

A correção da semântica das palavras, mesmo fora de ordem, é parte central da análise LDA, a qual é capaz de modelar a estrutura de cada reclamação através do teorema de Finetti (DIACONIS; FREEDMAN, 1980).

Após treinar o modelo, a persistência desse é atingida com o uso do módulo *Pickle*[7], o qual permite salvar estruturas de objeto Python na memória comum da máquina para uso posterior e recorrente. Assim, é possível salvar apenas o modelo para que se possa o reutilizar de maneira estática e rápida.

3.5. Visualização de resultados

A principal ferramenta utilizada para interpretação dos tópicos gerados foi a biblioteca pyLDAvis[i], a qual é uma conversão (também de código aberto) do pacote original (escrito para análise em R, e desenvolvido em R e D3) LDAvis[ii] (SIEVERT; SHIRLEY, 2014). A motivação inicial do desenvolvimento da ferramenta LDAvis se deu devido à dificuldade de interpretação dos tópicos gerados em uma análise LDA, em especial quando são somente considerados parâmetros de probabilidade de ocorrência. A ferramenta original é detalhada em *LDAvis: A method for visualizing and interpreting topics*[iii].

O objetivo da ferramenta LDAvis é possibilitar responder três perguntas: (1) *Qual o significado de cada tópico?* (2) *Quão prevalecente é cada tópico?* e (3) *Como os tópicos se relacionam?.* Uma visão geral da ferramenta pode ser vista na Figura 5.

[6] Disponível em https://radimrehurek.com/gensim/ . Licença GNU LGPLv2.1 .
[7] Mais em : https://docs.python.org/2/library/pickle.html . Obs: o uso de *Pickles* deve ser feito apenas em fase de desenvolvimento e teste, ou quando se tem certeza da procedência de seu conteúdo. O uso em eventual ambiente de produção deve ser evitado, como boa prática.

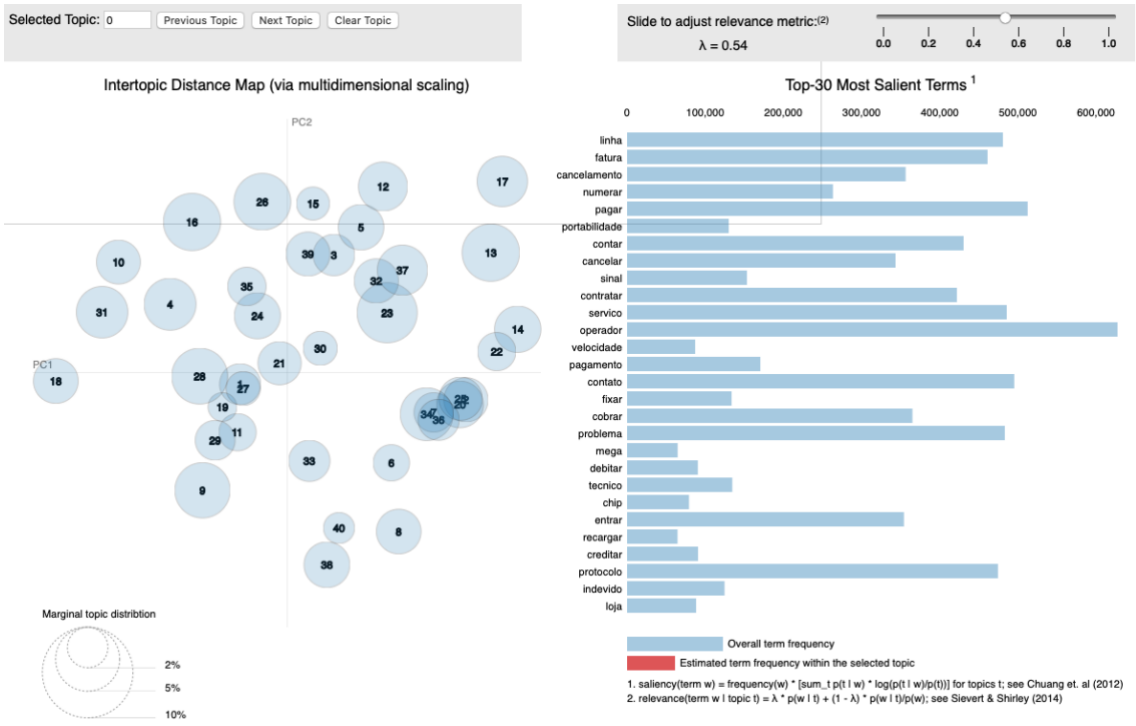


Figura 5 - Exemplo LDAvis em pyLDAvis.

O lado esquerdo da visualização permite que possamos ver a distribuição percentual de cada tópico ao longo dos textos de reclamação (os quais somam 100%). Como os tópicos não são exclusivos, mas podem se sobrepor em uma mesma reclamação, esse percentual é atingido com a razão entre o número de ocorrências de um determinado tópico e a soma de ocorrências de todos os tópicos. Um referencial de percentual é visível na parte inferior esquerda.

Ao selecionar um tópico (Figura 6), o percentual de prevalência/distribuição é destacado em texto na parte direita superior da análise. Além disso, as palavras mais importantes/relevantes para este tópico são destacadas abaixo da informação de percentual.

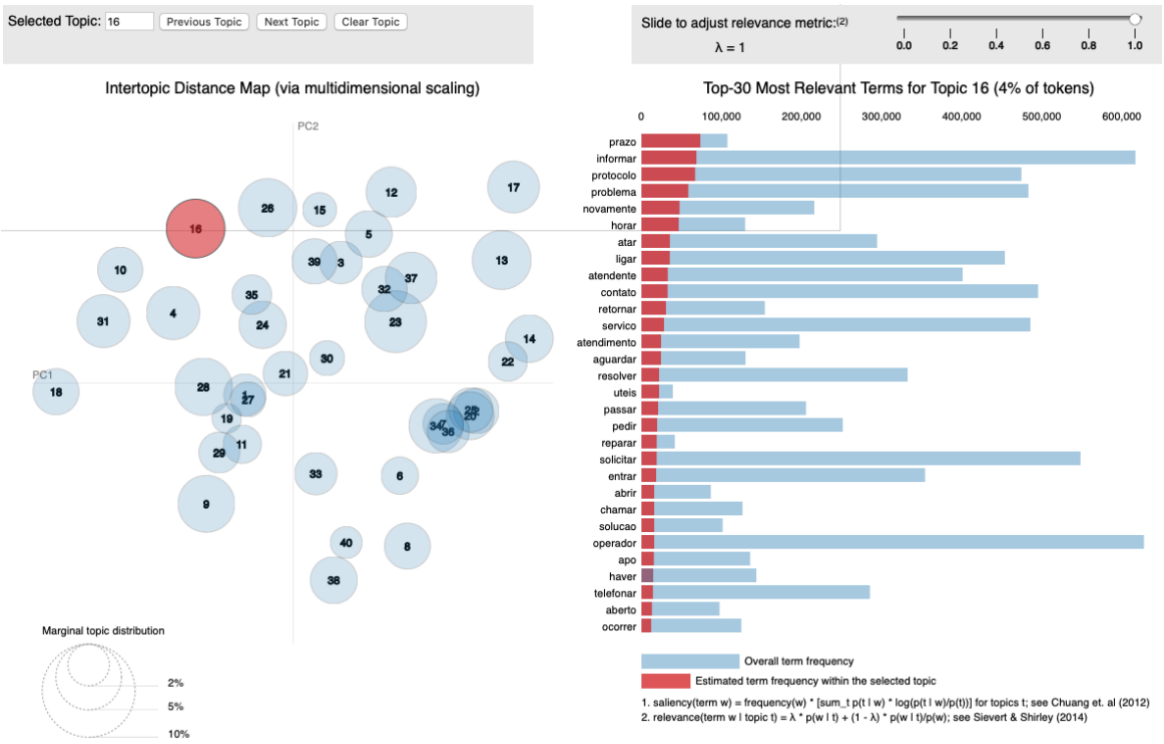


Figura 6 - Selecionar tópico em pyLDAvis

As barras azuis representam a frequência global de uma palavra no modelo, e a barra sobreposta vermelha representa a frequência específica da palavra no tópico. No entanto, a ordem com que as palavras aparecem quando seleciona-se um tópico considera o cálculo da **relevância de palavras por tópicos** (Equação 4) e por isso, ao alterar o parâmetro λ, é possível analisar palavras mais específicas de um tópico (0) ou de maior importância global (1). O cálculo da relevância pondera a probabilidade de aparecimento de uma palavra pela probabilidade de aparecimento da palavra em qualquer tópico.

Assim, a definição de relevância é:

Equação 4 - Definição relevância na análise LDAvis

$$relevância(palavra\ w|topic\ t) = \lambda * p(w|t) + (1-\lambda) * \frac{p(w|t)}{p(w)}$$

E quando $\lambda = 1$*:*

Equação 5 - relevância quando lambda = 1

$$relevância(palavra\ w|topic\ t, \lambda = 1) = 1 * p(w|t) + (1-1) * \frac{p(w|t)}{p(w)}$$

O que resulta em um simples cálculo da probabilidade de aparecimento de uma palavra w, dado um tópico t já escolhido, quando usado $\lambda = 1$.

Equação 6 - relevância quando lambda = 1, simplificado

$$relevância(palavra\ w|topic\ t, \lambda = 1) = p(w|t)$$

Por isso é importante o uso de diferentes parâmetros ao longo da interpretação de um tópico, a fim de possibilitar uma análise humana qualitativa das palavras que mais possam trazer poder explicativo do tópico.

Nesta análise, a **saliência de uma palavra nos tópicos** é calculada aplicando uma análise de componente principal, a qual pondera a probabilidade de uma palavra (w) aparecer em um contexto geral, pela soma da probabilidade de todos os contextos (tópicos) em que essa palavra pode aparecer (corrigido aplicando uma função log para suavizar os casos extremos a direita) (CHUANG; MANNING; HEER, 2012).

Assim, o cálculo da *distinção* de uma palavra é

Equação 7 - distinção

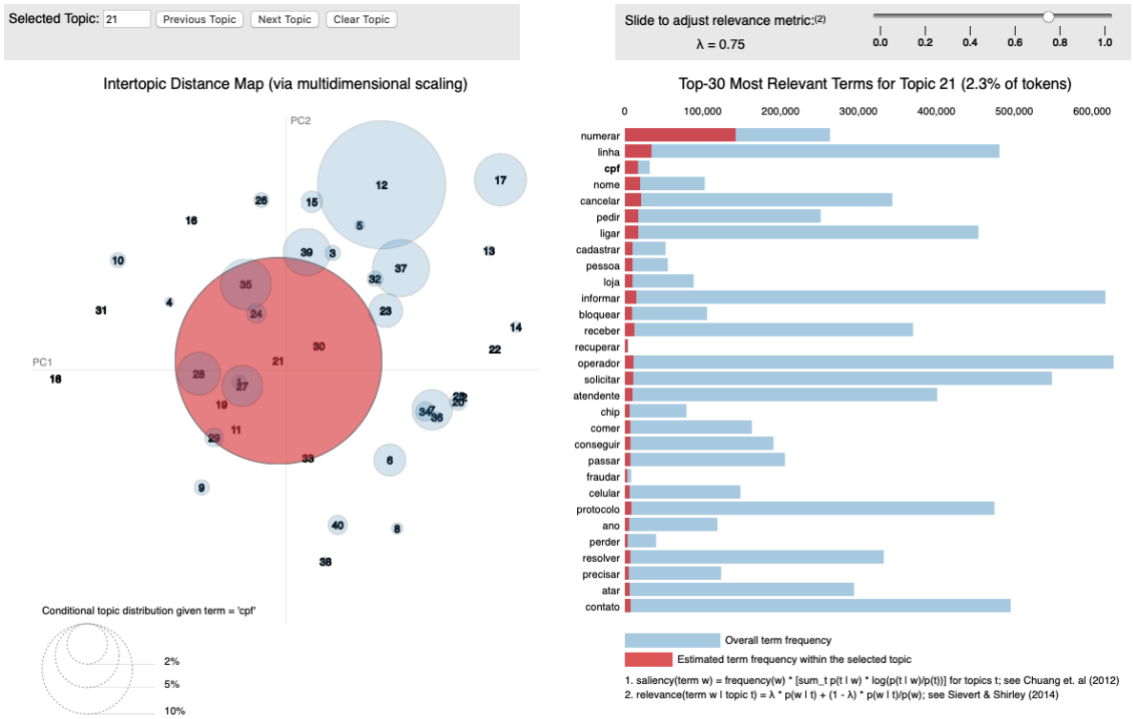
$$distinção(palavra\ w) = \sum_{T} P(T|w) * log \frac{P(T|w)}{P(T)}$$

E a saliência é

Equação 8 - saliência

$$saliência(palavra\ w) = P(w) * distinção(w)$$

Isso permite que, ao selecionar uma palavra do lado direito, seja possível ver sua saliência ao longo dos tópicos do lado esquerdo.



Para nossa análise, além dos resultados gerais provenientes da ferramenta *As-Is,* analisou-se a prevalência dos tópicos de reclamação ao longo de dois sub-grupos de dados: *Canais de entrada* e *Categorias de reclamação* (de acordo com classificação da Anatel). Como o objetivo não é criar um modelo para cada filtro, mas sim melhorar a interpretação em cima de um mesmo modelo, a alteração realizada altera somente a prevalência (e nenhum outro parâmetro) quando selecionada, fazendo com que as características do modelo treinado não se alterem.

A visualização dos resultados, para efeitos de compartilhamento com a equipe de análise do projeto e membros envolvidos da Anatel se dá pelo uso de plataforma web, a qual permite não só o uso das ferramentas com todas as funcionalidades, como também permite atualizações e melhorias iterativas. A versão atual do relatório técnico preliminar (a versão final, após o fim das análises do RT2 poderá ser disponibilizada no site do projeto, de acordo com interesse da Anatel) está disponível em forma de qr code (Figura 7 - Qr code plataforma análise NLP). [8]

Figura 7 - Qr code plataforma análise NLP

A plataforma do projeto é escrita em html, e hospedada no GitLab, junto aos códigos do projeto (privados). O site é acessível por qualquer pessoa com o link, e embora informações numéricas e e-mails tenham sido retirados, o compartilhamento do link deve ser direcionado apenas aqueles envolvidos com a análise, até que uma versão final seja produzida; validada e autorizada para indexação oficial em nome do projeto.

[8] Site disponível em https://projeto_anatel-ibict.gitlab.io/reclamacoesv2/index_v2_output.html

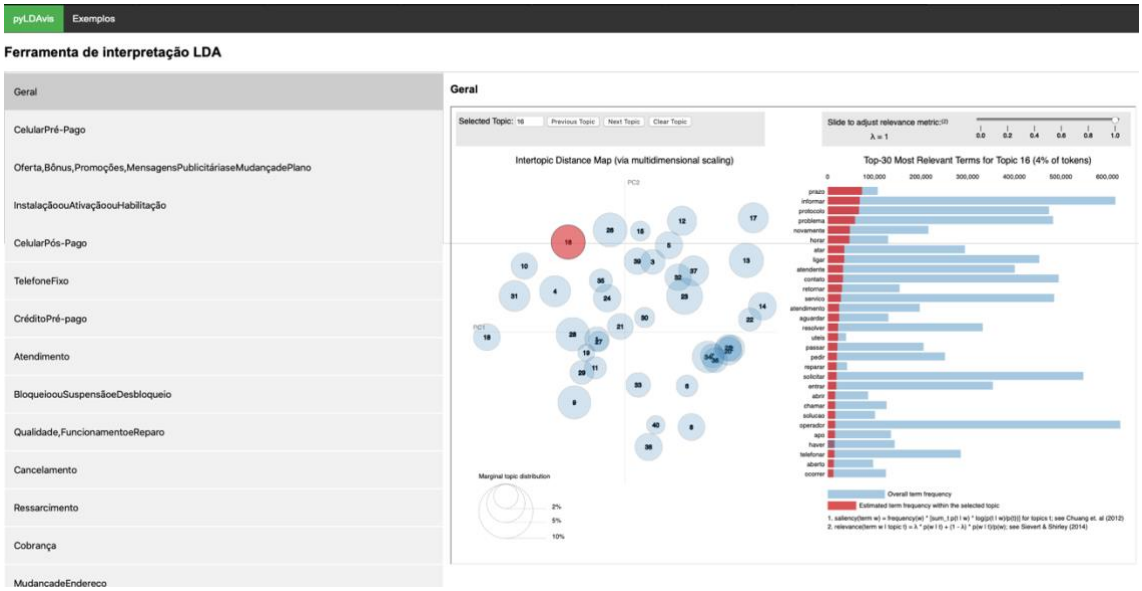


Figura 8 - Visualização plataforma análise NLP

A plataforma também possibilita a visualização de exemplos de reclamação, separados por tópicos. Ao selecionar o *header* Exemplos, é possível ler exemplos de reclamações que foram classificadas e tiveram o principal tópico como o selecionado (Figura 9 - exemplos de reclamações classificadas).

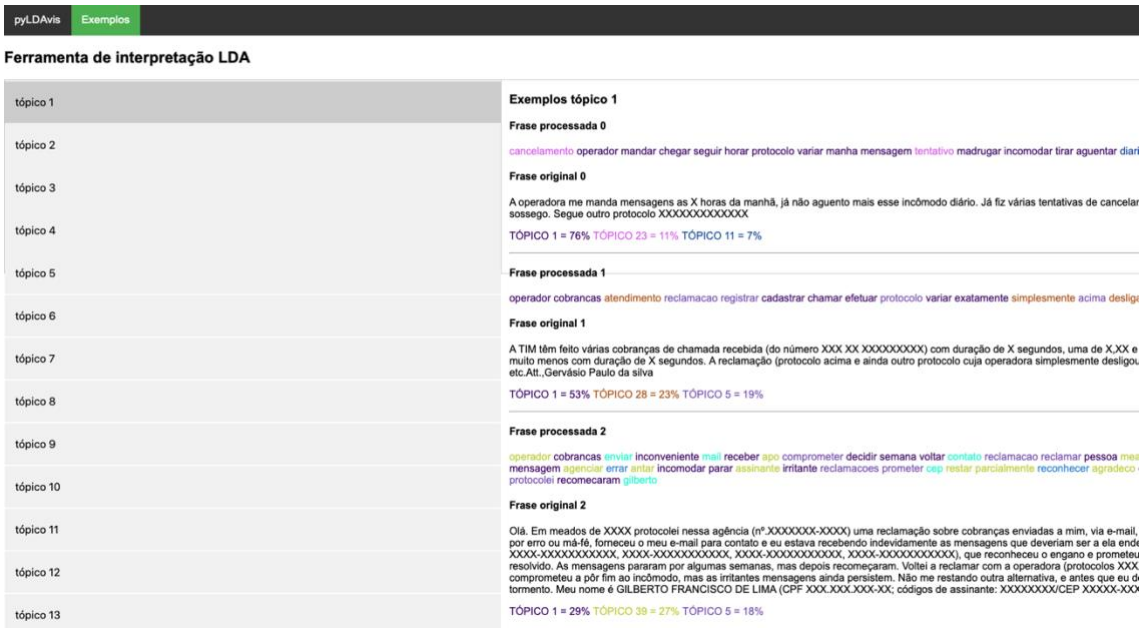


Figura 9 - exemplos de reclamações classificadas

Na análise exemplo de reclamações é possível conferir tanto o texto processado, como o texto original. O texto processado aparece colorido com as **cores referentes ao principal tópico que cada uma delas** participa (embora ela participe de outros além do que a cor indica). As cores são as mesmas usadas na classificação abaixo do exemplo original, as quais indicam o percentual que a reclamação específica tem de cada tópico. Como é possível perceber, as palavras não aparecem na mesma ordem em que são escritas, mas na mesma ordem em que são armazenadas no vetor de cada reclamação.

4. Interpretação preliminar dos tópicos

A partir da análise de mineração de texto com a técnica apresentada, foram gerados 40 tópicos que foram analisados, inicialmente, por dois pesquisadores distintos e de forma individual. Nesse primeiro momento, foi feita a leitura das palavras que cada um dos tópicos possui e também uma classificação de acordo com o entendimento acerca dos serviços de telecomunicações e temas de reclamação aos quais os tópicos possivelmente se referem. Vale ressaltar que na leitura de palavras os pesquisadores identificaram os termos de maior relevância para o tópico, sob diversos parâmetros possíveis (de termos mais específicos para aquele tópico até termos de relevância geral da classificação, ou seja, o parâmetro *lambda* entre 0 e 1 foi verificado 0.1, 0.2, 0.3….0.8, 0.9, 1.0). Após, foram feitas reuniões entre os dois pesquisadores mencionados e um terceiro pesquisador como mediador da discussão para que houvesse a comparação das duas análises prévias e, então, utilizando a técnica de validação por consenso, foi feita a formulação de uma nomenclatura final para cada tópico.

Em seguida, a análise supracitada foi apresentada a outros dois pesquisadores, que sugeriram a inclusão de determinadas análises, como: a etapa da jornada do consumidor que cada tópico se refere (Pesquisa/Prospecção, Contratação, Utilização, Pagamento, Engajamento), a análise funcional (aumento de punição ou redução do reforço) de cada tópico, a etapa de consumo (pré-compra, compra ou pós-compra) e também o tipo de problema (problema original que levou a reclamação ou problemas na cadeia de resposta de resolver - ou problemas derivados). Além disso, houve também a sugestão de agrupamento dos tópicos, em caso de semelhança. As sugestões foram aceitas e realizadas, o que gerou a seguinte tabela:

Grupo	Tópico	Nomenclatura FINAL	Etapa Jornada do Consumidor	OM - observação da função em vigor	Etapa de consumo	Tipo de problema
Reclamações acerca de Telemarketing	1	Incômodo com ligações de Telemarketing	Pesquisa/ Prospecção	Aumento da Punição	Pré e/ ou pós-compra	Problema original que levou a reclamação
Reclamações acerca de Ofertas	2	Divergência do contratado x entregue em ofertas em geral.	Utilização, Pagamento ou Engajamento	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
	8	Divergência do contratado x entregue em ofertas de TV.	Utilização, Pagamento ou Engajamento	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Instalação/ Reparação/ Visita Técnica	3	Problemas na instalação.	Contratação ou utilização	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
	4	Problemas com agendamento de	Contratação ou utilização	Aumento da Punição	Pós-compra	Problemas na cadeia de

		visita técnica em geral.				resposta de resolver (ou problemas derivados)
	16	Problemas com prazo de reparação em geral.	Utilização	Aumento da Punição	Pós-compra	Problemas na cadeia de resposta de resolver (ou problemas derivados)
	10	Problemas na instalação/visita técnica de serviços de banda larga.	Contratação ou utilização	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Reaberturas	5	Reclamação recorrente (reabertura)	Contratação ou utilização	Aumento da Punição	Pós-compra	Problemas na cadeia de resposta de resolver (ou problemas derivados)
Reclamações acerca de Planos	6	Problemas no cancelamento/habilitação de planos.	Engajamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
	19	Problemas com a renovação de bônus em geral.	Utilização ou Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
	20	Problemas com alteração de planos.	Utilização ou Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
	38	Alteração de planos sem autorização.	Utilização ou Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Cobranças indevidas	7	Cobrança de excedente na linha contratada por empresas.	Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
	17	Cobrança indevida e solicitação de ressarcimento em geral.	Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação

	22	Cobrança indevida (devido a débito automático e no cartão)	Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
	34	Cobrança indevida em faturas.	Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Atendimento	9	Insatisfação com o atendimento da operadora.	Engajamento	Aumento da Punição	Pós-compra	Problemas na cadeia de resposta de resolver (ou problemas derivados)
	28	Problemas com atendimento do prestador de serviço (canais).	Engajamento	Redução do Reforço	Pós-compra	Problemas na cadeia de resposta de resolver (ou problemas derivados)
Reclamações acerca de pré-pagos	11	Problemas com créditos pré-pagos.	Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
	33	Problemas com créditos pré-pagos.	Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Titularidade por Óbitos	12	Problemas com ajustes de titularidade por falecimento.	Engajamento	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Negociações	13	Problemas com acordo/ contestação de pagamentos de fatura.	Pagamento	Aumento da Punição	Pós-compra	Problemas na cadeia de resposta de resolver (ou problemas derivados)
	37	Problemas com negociação.	Pagamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de	14	Bloqueio/ suspensão	Utilização	Redução do Reforço	Pós-compra	Problema original que

Bloqueios indevidos		indevida de serviço.				levou a reclamação
Reclamações acerca de entrega de Aparelhos	15	Problemas com a retirada/ troca de aparelho.	Utilização	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Funcionamento técnico	18	Problemas com o sinal de celular.	Utilização	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
	24	Problemas técnicos de funcionamento	Utilização	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
	26	Problemas na prestação de serviços	Engajamento	Redução do Reforço	Pós-compra	Problemas na cadeia de resposta de resolver (ou problemas derivados)
	27	Problemas de acesso a aplicativos, sites e redes sociais.	Utilização	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
	30	Problemas com a velocidade da internet	Utilização	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
	31	Problemas técnicos com a Internet	Utilização	Redução do Reforço	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Fraudes	21	Problemas de fraude com celular.	Utilização, Pagamento ou Engajamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Cancelamentos	23	Problemas com solicitações de cancelamento em geral.	Engajamento	Aumento da Punição	Pós-compra	Problemas na cadeia de resposta de resolver (ou problemas derivados)
Reclamações acerca de Reajustes	25	Problemas com aumento nos	Pagamento	Aumento da Punição	Pós-compra	Problema original que

		valores cobrados (contrato)				levou a reclamação
Reclamações acerca de Vendas	29	Problemas com a compra/ troca de chip/ aparelho.	Contratação	Redução do Reforço	Compra	Problema original que levou a reclamação
	40	Problemas com vendas.	Pesquisa/Prospecção ou Contratação	Aumento da Punição	Pré-compra ou Compra	Problema original que levou a reclamação
Reclamações acerca de Contratos corporativos	32	Problemas com contratos corporativos.	Utilização, Pagamento ou Engajamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
Reclamações acerca de Portabilidades	35	Problemas com portabilidade.	Utilização ou Engajamento	Aumento da Punição	Pós-compra	Problema original que levou a reclamação
Dúvidas	36	VERIFICAR UTILIDADE DO TÓPICO (seria sobre a OI?)				
	39	VERIFICAR UTILIDADE DO TÓPICO - termos jurídicos				

Em seguida, os dados foram apresentados para a ANATEL em uma reunião que contou com os pesquisadores responsáveis e especialistas no assunto que atuam diretamente com as reclamações dos consumidores. Observou-se a necessidade de alguns exemplos de textos para cada tópico para que esses especialistas pudessem também analisar tópico por tópico. Os próximos passos envolvem uma nova reunião para validação por consenso entre os pesquisadores e os especialistas e, em seguida, a construção de um formulário para que haja uma validação de juízes da análise feita até então. Serão selecionados 3 outros especialistas que não participaram da discussão para essa tarefa e responderão individualmente o formulário, que contará com uma série de exemplos de reclamações e campos para indicar a quais tópicos se referem, com determinado nível de aderência.

5. Próximos passos

Este relatório, entregue em caráter preliminar, terá em sua versão final o desenvolvimento da atividade de identificação/interpretação dos tópicos individuais, bem como um maior detalhamento dos processos de investigação e testes utilizados durante o desenvolvimento do escopo deste relatório (e.g. Análises de cluster com k-means, regressões lineares, testes de análise de sensibilidade e testes de consistência de resultados matemáticos. Discussões sobre ajustes e refinamentos do modelo ao longo do processo de desenvolvimento também são serão apresentados, bem como possíveis aplicações e cenários no contexto da Anatel.

6. Referências

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

CHUANG, J.; MANNING, C. D.; HEER, J. **Termite: Visualization techniques for assessing textual topic models**. Proceedings of the international working conference on advanced visual interfaces. **Anais**...2012

DIACONIS, P.; FREEDMAN, D. de Finetti's theorem for Markov chains. **The Annals of Probability**, p. 115–130, 1980.

EXPLOSION, A. I. spaCy-Industrial-strength Natural Language Processing in Python. **URL: https://spacy. io**, 2017.

HOFFMAN, M.; BACH, F. R.; BLEI, D. M. **Online learning for latent dirichlet allocation**. advances in neural information processing systems. **Anais**...2010

REHUREK, R.; SOJKA, P. Gensim–python framework for vector space modelling. **NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic**, v. 3, n. 2, 2011.

SIEVERT, C.; SHIRLEY, K. **LDAvis: A method for visualizing and interpreting topics**. Proceedings of the workshop on interactive language learning, visualization, and interfaces. **Anais**...2014

---

[i] Código disponível em https://github.com/bmabey/pyLDAvis . Licença BSD-3.

[ii] Código e pacote disponível em https://cran.r-project.org/web/packages/LDAvis/index.html. Licença MIT + Propriedade intelectual AT&T.

[iii] Disponível em https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf .