Automatização do Processamento do Texto Bruto Oriundo de um Serviço de Atendimento de Reclamações

Maxwel de S. Freitas[1], Rodrigo V. Andreão[1]

[1]Programa de Pós-Graduação em Tecnologias Sustentáveis (PPGTECS) – Instituto Federal do Espírito Santo (IFES)
Avenida Vitória, 1729 – Jucutuquara – 29040-780 – Vitória – ES – Brazil

maxwelfreitas@anatel.gov.br, rodrigo.varejao@ifes.edu.br

Abstract. *The analysis of texts from customer service is an important tool for evaluating the quality of customer service provided to consumers by suppliers. In this paper we performed an exploratory analysis of the text extracted from the responses to complaints from consumers of telecommunications services to develop a cleaning routine capable of reducing the dimensionality and noise of the data representation model. The routine presented satisfactory results, reducing the dimensionality and noise of the data representation model, contributing to the construction of more efficient classifiers.*

Resumo. *Análise de textos oriundos de serviços de atendimento ao consumidor é importante ferramenta para avaliação da qualidade do atendimento dispensado aos consumidores pelos fornecedores. Neste trabalho foi realizada uma análise exploratória do texto extraído das respostas às reclamações de consumidores de serviços de telecomunicações para elaboração de uma rotina de limpeza capaz de reduzir a dimensionalidade e o ruído do modelo de representação de dados. A rotina apresentou resultados bastante satisfatórios, reduzindo a dimensionalidade e o ruído do modelo de representação de dados, contribuindo para a construção de classificadores mais eficientes.*

1. Introdução

A transformação digital ocorrida nos últimos anos proporcionou uma revolução nas comunicações humanas. A internet transformou-se no principal meio de comunicação entre as pessoas e impactou a forma como os consumidores se relacionam com seus fornecedores e prestadores de serviços. Cada vez mais exigentes, os consumidores têm acesso a inúmeros serviços públicos e privados, comércio eletrônico, plataformas de educação, atividades de lazer, e outros. [Anatel 2020b; ONU 2015].

Os consumidores apresentam suas demandas através de vários canais além dos serviços de atendimento ao consumidor: serviços de mensagens de texto, redes sociais ou sites especializados. As crescentes demandas geram um grande volume de dados não estruturados que devem ser capturados, analisados e tratados pelas empresas.

Para o setor de telecomunicações o desafio é maior: com 315,1 milhões de clientes, as prestadoras receberam em 2020 2,96 milhões de reclamações através da plataforma Anatel Consumidor, da Agência Nacional de Telecomunicações (Anatel) [Anatel 2020a, 2021].

Nos últimos anos, a Anatel promoveu diversas ações visando aprimorar a qualidade da prestação dos serviços, contudo, ainda persiste a necessidade de avanços a fim de promover a ampliação do acesso aos serviços de telecomunicações e a satisfação do consumidor, cada vez mais exigente quanto à qualidade desejada e dependente de telecomunicações para a realização de atividades cotidianas [Anatel 2020b]. Nesse contexto a Anatel instituiu o processo de Avaliação Qualitativa da Resposta para monitorar e aferir a qualidade do tratamento das prestadoras de serviços de telecomunicações às demandas de seus consumidores [Anatel 2020c].

No modelo de Avaliação Qualitativa da Resposta são avaliadas mensalmente uma amostra de respostas a reclamações de consumidores registradas pelas prestadoras dos cinco maiores grupos econômicos, as quais respondem por 92,5% dos acessos dos serviços de telefonia fixa, telefonia móvel, banda larga fixa e TV por assinatura no Brasil [Anatel 2020a]. Especialistas avaliam as respostas apresentadas pelas prestadoras e indicam se a resposta atendeu ou não ao item avaliado.

A grande quantidade de reclamações recebidas por mês, aliada ao tempo necessário para análise de cada resposta e a pouca quantidade de servidores dedicados ao processo, são fatores impeditivos para uma análise manual mais ampla. Uma abordagem de processamento de linguagem natural pode ser utilizada para ampliar o alcance da Avaliação Qualitativa da Resposta e analisar automaticamente todas as reclamações registradas, aumentando a eficiência do processo e melhorando a qualidade do gasto público.

Neste contexto, o objetivo desse trabalho é automatizar o processo de tratamento de texto bruto oriundo de um serviço de atendimento de reclamações tendo em vista a sua aplicação futura na construção de métodos de aprendizado de máquina e classificação de textos. Para isso, é realizada uma análise exploratória das respostas às reclamações de consumidores avaliadas no processo de Avaliação Qualitativa da Resposta. A metodologia proposta busca por um lado reduzir a dimensionalidade dos atributos de entrada, e por outro reduzir o ruído dos dados, beneficiando a sua aplicação futura na construção de classificadores automáticos.

2. Revisão de literatura

Técnicas de processamento de linguagem natural são utilizadas em várias aplicações em diversos domínios. Análise de textos não estruturados utilizando técnicas de mineração de texto e de dados podem ajudar bastante os processos de tomada de decisão. Aplicações típicas de processamento de linguagem natural incluem sumarização de documentos [García Adeva et al. 2014], classificação e priorização de mensagens [Gomez and Moens 2012; Sulieman et al. 2017], análise de sentimentos [Silva, Bonfante e Martins, 2017], detecção de mensagens eletrônicas indesejadas [Guzella and Caminhas 2009; Liu et al. 2016], análise de reclamações de consumidores [Faed et al. 2016; Ordenes et al. 2014], classificação de patentes [Mascarenhas e Bonfante, 2017], entre outras.

Classificação de texto, ou categorização de texto, é a técnica empregada para analisar documentos e a eles atribuir categorias previamente definidas. A Figura 1 ilustra um sistema típico de classificação de texto por aprendizagem de máquina supervisionada. Na fase de treinamento, os documentos classificados manualmente são transformados de texto bruto para um formato adequado para o algoritmo de aprendizagem de máquina, a partir do qual um classificador é construído [Aas and Eikvil 1999]. Uma vez treinado, o

sistema é capaz de analisar novos documentos, classificando-os de forma similar a um especialista humano [García Adeva et al. 2014], permitindo a análise e classificação de um universo de documentos ao invés de uma pequena amostra.

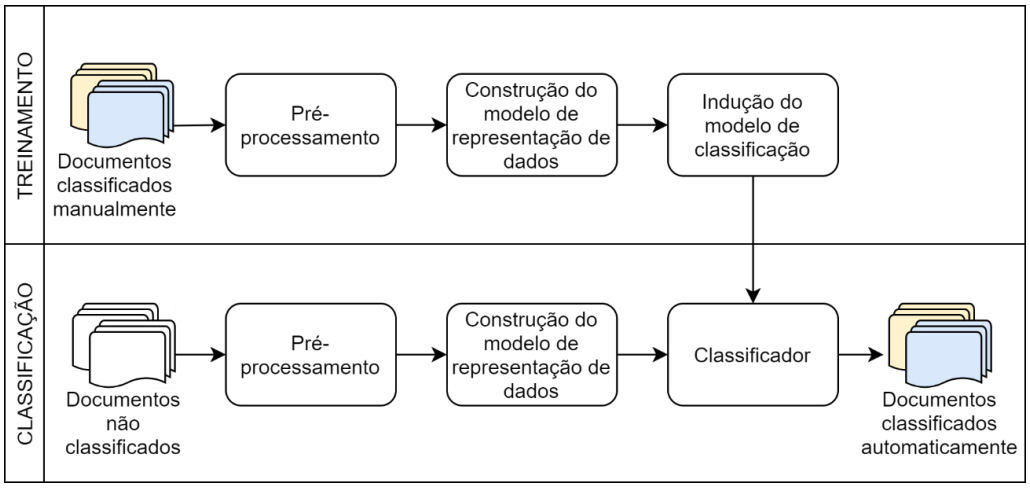

Figura 1. Ilustração de um sistema típico de classificação de texto por aprendizagem de máquina supervisionada

Na etapa de pré-processamento, os documentos são coletados e submetidos a procedimentos de limpeza e preparação do texto para que sejam mantidas apenas as palavras semanticamente relevantes. São removidas as marcações do texto, caracteres não textuais, números, stop words (pronomes, preposições, conjunções, etc.) e as palavras são reduzidas ao seu tronco ou à sua forma canônica [Aas and Eikvil 1999; Aggarwal 2015].

Após o pré-processamento, os documentos são convertidos em uma representação adequada para o algoritmo de aprendizagem de máquina. Uma das abordagens mais comuns é o modelo de espaço vetorial, também conhecido como *bag-of-words*, no qual cada documento é representado por um vetor de termos, e o conjunto de documentos é representado por uma matriz termo-documento na qual cada elemento representa o peso de cada termo em um documento [Aas and Eikvil 1999; Salton et al. 1975].

A alta dimensionalidade e a esparsidade do espaço vetorial são um problema central em classificação de textos baseada em estatística. Uma única palavra que esteja presente em apenas um documento corresponde a uma linha da matriz termo-documento. Existindo vários documentos na coleção, a quantidade de linhas pode facilmente chegar a centenas de milhares [Aas and Eikvil 1999].

Após a preparação do conjunto de documentos e sua transformação no modelo de espaço vetorial, métodos ou algoritmos de aprendizagem de máquina são aplicados para induzir uma função de classificação para mapear os documentos a uma classe. Uma vez induzida, a função de classificação é aplicada aos documentos não classificados, cujas classes são desconhecidas [Manning et al. 2008].

O desempenho dos modelos de classificação construídos pode ser avaliado com base na quantidade de documentos classificados corretamente. A avaliação de modelos de classificação atende a vários objetivos: comparar diferentes modelos de classificadores, selecionar o melhor classificador para um conjunto de dados específico ou ajustar os parâmetros do classificador. A avaliação de desempenho é feita aplicando-se vários

algoritmos a um ou mais subconjunto de documentos e classificando os resultados obtidos [Aggarwal 2015; Sokolova and Lapalme 2007].

3. Metodologia

A metodologia adotada nesse trabalho consistiu, basicamente, nas etapas de seleção da base de dados e pré-processamento de texto, na qual o texto bruto das respostas às reclamações dos consumidores previamente avaliadas foi extraído da plataforma Anatel Consumidor e preparado para a construção do modelo de representação de dados.

3.1. Seleção da Base de Dados

As fontes primárias de dados para a pesquisa são as planilhas de Avaliação Qualitativa da Resposta (AQR), elaboradas mensalmente com dados extraídos da plataforma Anatel Consumidor. Cada planilha contém uma amostra de 384 reclamações dos cinco maiores grupos econômicos: Claro, Oi, Tim, Vivo e Sky.

Cada registro (linha) da planilha AQR possui 61 campos (colunas) divididos em dois grupos: 17 campos descritivos e 44 campos avaliativos. Os campos avaliativos são divididos, ainda, em 4 subgrupos: assuntos, anormalidades, itens avaliados e comentários. Os campos descritivos são extraídos diretamente da plataforma Anatel Consumidor e os campos avaliativos são preenchidos pelo especialista humano (avaliador).

Para desenvolvimento deste trabalho foi construída uma tabela de respostas avaliadas com 34 campos obtidos a partir das planilhas AQR: o identificador único de cada reclamação (protocolo), o código do assunto da reclamação e os 33 itens avaliados pelos especialistas humanos (avaliador), além do histórico de respostas da prestadora às reclamações, obtido em consulta ao banco de dados da plataforma Anatel Consumidor, usando o número do protocolo como chave de pesquisa.

Neste trabalho foram utilizadas as planilhas AQR elaboradas no período de junho a dezembro de 2020, contendo um total de 13.440 reclamações avaliadas.

3.2. Pré-Processamento da Base de Dados

A partir das planilhas AQR foram relacionados os protocolos de cada reclamação e consultados, diretamente na base de dados da plataforma Anatel Consumidor, os históricos das respostas apresentadas pelas prestadoras. Foram consideradas apenas as interações marcadas como resposta. Das 13.440 reclamações foram obtidos 13.434 documentos, pois três reclamações estavam duplicadas na amostra e outras três estavam com resposta indicada incorretamente.

Para subsidiar a análise foram obtidos ainda os nomes e sobrenomes de consumidores registrados na plataforma Anatel Consumidor e os verbetes do Vocabulário Ortográfico da Língua Portuguesa (VOLP) [Bechara 2017].

Os documentos foram submetidos a rotinas de processamento utilizando a linguagem Python juntamente com as bibliotecas de processamento de linguagem natural spaCy[1] e de aprendizado de máquina Scikit-Learn [Pedregosa et al. 2011].

[1] https://spacy.io/

Os documentos foram convertidos para minúsculas e, em seguida, processados pelo módulo de segmentação do spaCy. Os documentos foram divididos em segmentos (do inglês *tokens*) e cada segmento foi classificado automaticamente em quatro grupos: caracteres de pontuação, palavras frequentes (do inglês *stopwords*), números e endereços eletrônicos.

Após o processamento pelo spaCy foram contadas as frequências de ocorrência dos segmentos únicos na coleção de documentos. Os segmentos que não foram classificados automaticamente pelo spaCy foram classificados em outros três grupos: grandes (com mais de 30 caracteres), nomes próprios e presentes em apenas um documento. Foi atribuído a cada segmento um único grupo, observando a ordem de prioridade: a) caracteres de pontuação ou palavras frequentes; b) números; c) endereços eletrônicos; d) grandes; e) nomes próprios; f) uma ocorrência na coleção de documentos e g) outros.

Os segmentos classificados em um dos primeiros seis grupos são considerados ruído, pois não possuem informação semanticamente relevante para o domínio em estudo, logo, podem ser descartados no processo de construção do modelo de espaço vetorial. A distribuição do tamanho dos demais segmentos foi comparada visualmente com a distribuição do tamanho dos verbetes do VOLP para identificação de padrões e tipos de segmentos não identificados adequadamente. A partir das divergências identificadas na análise visual dos histogramas foi construída uma rotina para pré-processamento do texto previamente à segmentação.

Para aferir o impacto do pré-processamento do texto bruto em tarefas de classificação de texto foi foram construídos três modelos de espaço vetorial: com texto bruto, sem qualquer tratamento prévio; com texto sem caracteres não alfanuméricos e palavras frequentes; e com texto processado pela rotina desenvolvida.

Os documentos foram divididos em duas classes: 1) a prestadora informou contato com sucesso com o reclamante ou 2) a prestadora não informou contato com sucesso com o reclamante. Cada modelo foi submetido a sete algoritmos de aprendizagem de máquina e foram apuradas a precisão obtida e o tempo de aprendizado de cada modelo.

4. Resultados

Na primeira segmentação, executada no sem qualquer pré-processamento do texto, foram obtidos 92.896 segmentos únicos, classificados em sete grupos, ilustrados na Figura 2. A grande maioria dos segmentos são números ou estão presentes em apenas um documento da coleção, situação já esperada, uma vez que cada resposta faz referência a muitos números únicos para cada uma delas (protocolos, telefones, contratos).

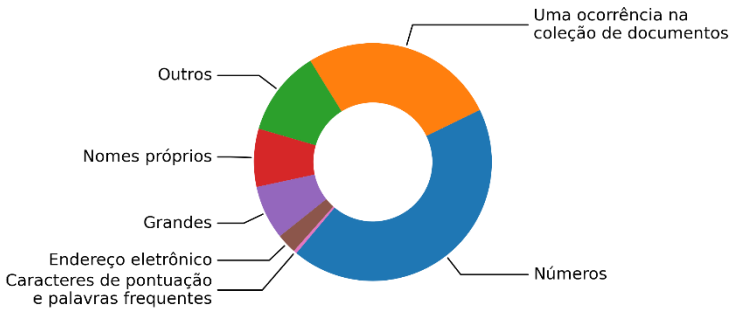

Figura 2. Distribuição de segmentos obtidos sem pré-processamento do texto

Os segmentos obtidos a partir do texto bruto (Figura 3b) e do texto submetido a limpeza simples, com remoção de caracteres não alfanuméricos e palavras frequentes (Figura 3c) apresentaram distribuição de tamanho visualmente diferentes da distribuição de tamanho dos verbetes do VOLP (Figura 3a).

A grande quantidade de segmentos de texto bruto com tamanho 5, 8, 10, 16 ou mais de 20 é decorrente de números com caracteres não numéricos (números de telefone ou contrato separados por hífen, datas e horas) e partes de identificadores de assinaturas que não foram adequadamente identificados e tratados pelo segmentador. Tais segmentos são considerados ruído, pois não são significativos para o que se presente avaliar, além disso sua presença resulta em um modelo de representação de dados com dimensionalidade maior que a adequada.

A partir da observação e análise dos resultados da primeira segmentação foi elaborada uma rotina para pré-processamento do texto para identificar e remover todos os termos que não foram adequadamente identificados e tratados pelo segmentador, cujo resultado foi um conjunto de segmentos com distribuição de tamanho (Figura 3d) muito próxima da distribuição de tamanho dos verbetes do VOLP (Figura 3a).

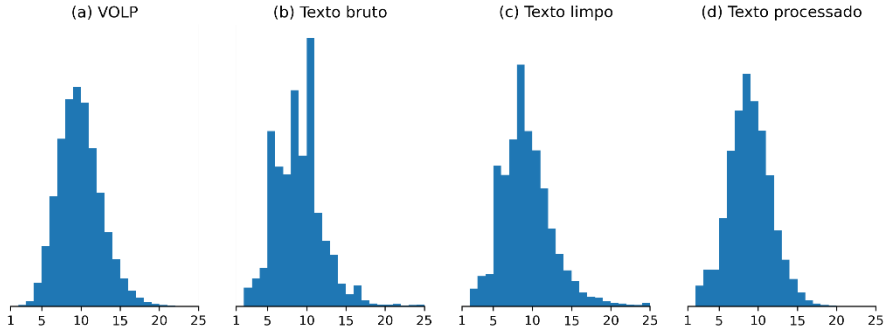

Figura 3. Comparação do tamanho dos segmentos com o tamanho dos verbetes do VOLP

O modelo de espaço vetorial resultante do texto processado contém 6.508 atributos, uma dimensionalidade 49,9% menor que o resultante do texto bruto, com 12.990 atributos. Já o espaço vetorial resultante do texto limpo contém 14.795 atributos, 13,9% a mais que o do texto bruto.

Não houve variação significativa na acurácia obtida dos algoritmos de classificação para os diferentes modelos de espaço vetorial (Tabela 1), contudo, o tempo de treinamento foi significativamente menor no modelo de espaço vetorial construído a partir do texto processado (Tabela 2).

Tabela 1. Acurácia dos modelos de classificação

Classificador	Texto Bruto	Texto Limpo	Texto Processado
Árvore de decisão	80,28	79,51	77,70
K Vizinhos Mais Próximos	74,59	74,28	75,50
Regressão Logística	81,11	80,82	81,62
Naive Bayes	75,20	74,77	75,07
Floresta Aleatória	80,27	80,46	81,15
Regressão Ridge	82,20	80,92	82,13
Máquina de Vetores de Suporte	82,24	81,60	82,38

Tabela 2. Tempo de treinamento dos modelos de classificação (segundos)

Classificador	Texto Bruto	Texto Limpo	Texto Processado
Árvore de decisão	49,26	41,53	35,78
K Vizinhos Mais Próximos	95,56	65,02	56,94
Regressão Logística	6,38	5,65	2
Naive Bayes	0,26	0,24	0,16
Floresta Aleatória	72,58	68,43	60,79
Regressão Ridge	2,47	1,92	0,59
Máquina de Vetores de Suporte	1.150,53	986,56	592,52

5. Conclusão

Neste trabalho foi apresentada uma rotina de processamento de texto bruto capaz de reduzir a dimensionalidade e o ruído do modelo de representação de dados, contribuindo para a construção de classificadores de texto mais eficientes.

Os resultados obtidos apresentam os impactos de duas rotinas de processamento de texto no resultado de classificadores de texto, construídos com algoritmos de aprendizado de máquina supervisionada, em termos de eficácia e de tempo de treinamento. No domínio estudado a segmentação e análise do texto bruto permitiu obter informações mais precisas sobre o texto em análise e a construção de uma rotina de processamento adequada às suas características.

Como trabalhos futuros pretende-se construir outros modelos de representação de dados e de aprendizagem de máquina mais recentes na literatura para encontrar a solução mais adequada ao problema em estudo.

6. Referências

Aas, K. and Eikvil, L. (1999). Text Categorisation: A Survey.

Aggarwal, C. C. (2015). Data Mining. Cham: Springer International Publishing.

Anatel (2020a). Painéis de Dados da Anatel. https://www.anatel.gov.br/paineis/. Acessado em: 2 junho 2021.

Anatel (2020b). Plano Estratégico Conectando a Anatel ao Futuro. . https://www.anatel.gov.br/institucional/acoes-e-programas/planejamento-estrategico.

Anatel (2020c). Instruções - Planilha de Avaliação Qualitatia da Resposta. . https://sei.anatel.gov.br/sei/modulos/pesquisa/md_pesq_documento_consulta_externa.php?eEP-wqk1skrd8hSlk5Z3rN4EVg9uLJqrLYJw_9INcO7w9Wxkpaqnal8nvHu3R8amg6pefuNLP33rRWNm90bnp6LYqQxE9731uh2boPWGVST1NTgknbJs7-MMbFGD49DW. Acesso em: 11 fevereiro 2021.

Anatel (2021). Panorama - Reclamações 2020. . https://sistemas.anatel.gov.br/anexar-api/publico/anexos/download/fbbd24928ef71087dd56d8c8e9a99cab. Acesso em: 11 fevereiro 2021.

Bechara, E. [Ed.] (2017). Vocabulário Ortográfico da Língua Portuguesa. 6.a ed. Rio de Janeiro (RJ): Academia Brasileira de Letras.

Faed, A., Chang, E., Saberi, M., Hussain, O. K. and Azadeh, A. (2016). Intelligent customer complaint handling utilising principal component and data envelopment analysis (PDA). Applied Soft Computing Journal, v. 47, p. 614–630.

García Adeva, J. J., Pikatza Atxa, J. M., Ubeda Carrillo, M. and Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. Expert Systems with Applications, v. 41, n. 4 PART 1, p. 1498–1508.

Gomez, J. C. and Moens, M. F. (2012). PCA document reconstruction for email classification. Computational Statistics and Data Analysis, v. 56, n. 3, p. 741–751.

Guzella, T. S. and Caminhas, W. M. (2009). A review of machine learning approaches to Spam filtering. Expert Systems with Applications, v. 36, n. 7, p. 10206–10222.

Liu, Y., Wang, Y., Feng, L. and Zhu, X. (2016). Term frequency combined hybrid feature selection method for spam filtering. Pattern Analysis and Applications, v. 19, n. 2, p. 369–383.

Manning, C. D., Raghavan, P. and Schütze, H. (2008). Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Mascarenhas, Tamara A. T., Bonfante, Andreia G. (2017). Classificação de documentos de patentes usando o Doc2vec. Anais da VI Escola Regional de Informática da Sociedade Brasileira de Computação (SBC)-Regional de Mato Grosso, p. 28.

ONU (2015). Transformando Nosso Mundo: A Agenda 2030 para o Desenvolvimento Sustentável. https://nacoesunidas.org/wp-content/uploads/2015/10/agenda2030-pt-br.pdf. Acessado em: 2 junho 2021.

Ordenes, F. V., Theodoulidis, B., Burton, J., Gruber, T. and Zaki, M. (2014). Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach. Journal of Service Research, v. 17, n. 3, p. 278–295.

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. JMLR. v. 12, n. 85, p. 2825-2830.

Salton, G., Wong, A. and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. Commun. ACM, v. 18, n. 11, p. 613–620.

Silva, Fábio R. A., Bonfante, Andreia G., Martins, Claudia A. (2017). Deep Learning Aplicado à Análise de Sentimento em Linguagem Figurativa. Anais da VI Escola Regional de Informática da Sociedade Brasileira de Computação (SBC)-Regional de Mato Grosso, p. 12.

Sokolova, M. and Lapalme, G. (2007). Performance measures in classification of human communications. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v. 4509 LNAI, p. 159–170.

Sulieman, L., Gilmore, D., French, C., et al. (2017). Classifying patient portal messages using Convolutional Neural Networks. Journal of Biomedical Informatics, v. 74, p. 59–70.