Avaliação Automática de Conteúdo de Aplicações de Reclamação Online

Lucas G. S. Félix[1], João Victor Silveira[1], Washington Luiz[2], Diego Dias[1], Leonardo Rocha[1]

[1] Departamento de Ciência da Computação - Universidade Federal de São João del-Rei
Av. Visconde do Rio Preto S/N - Colônia do Bengo
lucasgsfelix@ufsj.edu.br, caetanosjoao@ufsj.edu.br,
diegodias@ufsj.edu.br, lcrocha@ufsj.edu.br
[2] Departamento de Ciência da Computação - Universidade Federal de Minas Gerais
washingtoncunha@dcc.ufmg.br

Abstract. A Internet tem vivenciado uma notória expansão e popularização nos últimos anos. Estima-se que até o ano de 2020 haja cerca de 40 trilhões de gigabytes de dados gerados. Existem diversos cenários onde novas técnicas e metodologias vêm sendo propostas para que informações relevantes possam ser extraídas desse grande volume de dados. Um exemplo recente são as aplicações reclamações online, tais como ReclameAqui, que funcionam como porta voz de consumidores insatisfeitos que relatam suas experiências ruins com determinados produtos e/ou serviços. Esses dados podem representar uma rica fonte de informação que pode ser utilizada por empresas em seu aperfeiçoamento. Nesse trabalho propomos uma metodologia que, por meio da combinação de técnicas de modelagem de tópicos e análise de sentimento, é capaz de extrair desses dados informações úteis, ricas em detalhes, que possam contribuir para empresas identificarem de forma mais consistente e rápida problemas nos produtos e serviços. Avaliamos nossa metodologia com coleção de comentários coletados da aplicação ReclameAqui, outra do Twitter e outra do PROCON, todas elas relacionadas às quatro maiores empresas de telefonia do Brasil (TIM, OI, VIVO e CLARO). Em nossas avaliações demonstramos que a riqueza de detalhes que podem ser extraídas do ReclameAqui e do Twitter são bem maiores quando comparadas a aquelas registradas no PROCON. Além disso, demonstrando que, por ser uma aplicação extremamente informal, extrair informações do Twitter exige mais recurso computacional e humano, o que torna os comentários de aplicações de reclamação online a melhor alternativa para se extrair informações úteis.

Categories and Subject Descriptors: H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing; H.2.8 [**DATABASE MANAGEMENT**]: Database Applications

Keywords: Modelagem de Tópicos, Análise de Sentimento, Aplicações Internet-Based

1. INTRODUÇÃO

A Internet tem vivenciado uma notória expansão e popularização nos últimos anos. Novos componentes e serviços estão sendo incorporados em um ritmo bastante acelerado e cada vez mais, novos usuários fazem uso desses serviços. A cada dia são criadas novas aplicações que geram e utilizam uma quantidade maior de dados dos mais diversos tipos e que atingem, senão a todos, quase todos os grupos de usuários. Estima-se que até o ano de 2020 haja cerca de 40 trilhões de gigabytes de dados gerados [1]. Esse grande volume de dados disponível na Web gerou nos últimos anos um diferenciado, desafiante e intrigante cenário para variadas aplicações: há mais dados que efetivamente pode-se analisar, como afirmado em [Auden 2002], "Muita informação é tão ruim quanto nenhuma.". Dessa forma, organizar e encontrar os recursos informacionais apropriados para satisfazer as necessidades dos usuários passou a figurar como um dos problemas mais desafiadores em Ciência da Computação.

[1] www.bigdatabusiness.com.br/os-grandes-e-impressioantes-numeros-de-big-data/

Diversos são os cenários onde novas técnicas e metodologias vêm sendo propostas para que informações relevantes possam ser extraídas. Um exemplo de aplicação interessante são as chamadas redes sociais (Twitter, Blogs, Facebook, etc.) nas quais as pessoas estão cada vez mais publicando suas opiniões na Web. Outro exemplo são as aplicações reclamações online, tais como ReclameAqui [2], Proteste [3] e Denuncio [4], que funcionam como porta voz de consumidores insatisfeitos que utilizam dessas aplicações para relatar problemas com produtos, empresas e serviços, etc. utilizando uma linguagem mais informal que os meios mais tradicionais como PROCON. Os dados disponibilizados por essas aplicações podem representar uma rica fonte de informação que pode ser utilizados por empresas para aperfeiçoar seus serviços, produtos, etc., e estabelecer uma relação mais próxima com seus clientes. Uma vez que a manipulação manual deste volume de dados é impraticável, temos, recentemente, a adaptação de áreas tradicionais de análise a este novo cenário, tais como Análise de Sentimento e Modelagem de Tópicos, sendo esses os temas do artigo aqui apresentado.

Nesse trabalho propomos uma metodologia que visa avaliar automaticamente os comentários realizados por usuários em sistemas de reclamação com o objetivo de extrair informações úteis, ricas em detalhes, que possam contribuir para empresas identificarem de forma mais consistente e rápida problemas nos produtos e serviços. Basicamente, a metodologia consiste em realizar coletas de dados referentes a alguma empresa específica, tratar adequadamente esses dados por meio de técnicas de preprocessamento de texto. A partir desses dados, o passo seguinte consiste em aplicar técnicas de modelagem de tópicos [Cheng et al. 2014] para inferir grupos de usuários e/ou comentários semanticamente relacionados em torno de alguma característica discutida. O objetivo é que a partir desses tópicos empresas sejam capazes de identificar reclamações relacionadas ao mesmo tema, apresentar uma solução para o mesmo de forma mais rápida, além de otimizar o processo de atendimento ao consumidor.

Para validar nossa metodologia, aplicamos a mesma em dados reais, coletados a partir da aplicação ReclameAqui, relacionados a comentários sobre as principais empresas de telecomunicação que atuam no Brasil (TIM, OI, CLARO e VIVO). A motivação para a escolha dessas empresas se dá pela grande quantidade de usuários que as mesmas possuem, sendo mais de 235 milhões de linhas móveis atualmente [5], a variedade de serviços apresentados pelas mesmas, tais como telefonia, internet móvel, TV a cabo, entre outros. Comparamos os resultados obtidos a partir desses dados com outras duas fontes de informação: reclamações registradas no PROCON e menções a essas empresas no Twitter. A primeira fonte é mais formal, com dados estruturados e hierarquicamente organizados, seguindo uma taxonomia previamente definida. Já os dados coletados a partir do twitter são completamente informais e vão desde propagandas sobre as empresas, até insultos com emprego de palavras de baixo calão. Para essa base, além do preprocessamento de texto, aplicamos também técnicas de análise de sentimento [Almeida et al. 2016] para filtrar quais eram efetivamente reclamações (sentimento negativo), além de removermos aqueles que continham palavras de baixo calão. Somente após essa etapa é que aplicamos a estratégia de modelagem de tópicos. O objetivo da comparação é avaliar a riqueza de detalhes que podem ser obtidos a partir dos comentários do ReclameAqui, comparando com um cenário mais formal e outro completamente informal. Nossas análises demonstraram que a partir dos comentários do ReclameAqui é possível identificar problemas relacionados a essas empresas de forma mais específica quando comparados aos mencionados na base de dados do PROCON, sem a poluição de informações observadas na coleção de dados do Twitter.

2. TRABALHOS RELACIONADOS

Conforme mencionado na Introdução, as duas principais técnicas de mineração de dados utilizadas em nosso trabalho estão relacionadas à analise de sentimento e modelagem de tópicos. Sendo assim,

[2] reclameaqui.com.br
[3] www.proteste.org.br
[4] denuncio.com.br
[5] www.anatel.gov.br/dados/destaque-1/283-brasiltem-236-2-milhoes-de-linhas-moveis-em-janeiro-de-2018

nesta seção apresentamos e discutimos alguns dos principais trabalhos dessas áreas.

2.1 Análise de sentimentos

Análise de sentimentos consiste em detectar automaticamente a polaridade (positiva, neutra, negativa) de um texto e vem sendo aplicada em diversas áreas para modelagem e entendimento do comportamento de usuários e análise de opiniões em textos [Rocha et al. 2015]. Atualmente, são considerados na literatura dois tipos de de métodos para análise de sentimento em texto: métodos baseados em aprendizado de máquina (AM) e métodos baseados em léxicos. Os trabalhos [Pak and Paroubek 2010; Almeida et al. 2016] utilizam-se de métodos de AM para classificação do texto. Os trabalhos correlatados se diferenciam apenas nos contextos de suas bases de dados e no algoritmo de classificação que é utilizado, de maneira geral possuindo poucas diferenças. No trabalho de [Pak and Paroubek 2010] é utilizado o algoritmo *Naive Bayes* para classificação dos dados e no artigo de [Almeida et al. 2016] são utilizados *Multinominal Naive Bayes, Support Vector Machine e Maximum Entropy* que são considerados estado da arte na classificação de texto. Vale destacar, entretanto, que essas soluções são sensíveis ao contexto para um dado conjunto de entrada, ou seja, para cada base de treinamento é necessário revalidar o classificador e criar-se um novo dicionário rotulando os dados.

Nos trabalhos de [Rocha et al. 2015; Sá et al. 2014; Gilbert 2014] são propostos métodos baseados em modelos léxicos para análise de sentimentos em textos. O artigo de [Rocha et al. 2015] propõe um método para análise de sentimento em texto utilizando uma técnica não-supervisionada que consegue extrair de grande volume de dados o sentimento coletivo sem realizar classificações individuais utilizando-se de um grafo probabilístico. Já o estudo de [Sá et al. 2014], propõe um léxico que trabalha através de um método semi-supervisionado ciente de contexto assumindo que existem diferentes classes de palavras que possuem diferentes comportamentos no significado de uma frase. A produção [Gilbert 2014] propõe um léxico para análise de sentimentos baseado em regras de outros métodos que trabalham sob o mesmo problema, sendo este considerado estado da arte nesta área de estudo.

2.2 Modelagem de tópicos

Modelagem de tópicos (MT) é um problema que envolve a descoberta de relações entre documentos e tópicos, assim como a descoberta de relações entre termos que compõe os documentos e os tópicos permitindo assim a organização dos documentos de uma coleção em tópicos semânticos [Luiz et al. 2018]. Atualmente, a modelagem de tópicos tem sido uma técnica amplamente utilizada para avaliação automática de enormes quantidades de texto. De maneira geral pode-se descrever essa prática através da sua capacidade de reduzir documentos e representa-los através de tópicos.

Nesse contexto, foram avaliadas produções que aplicam diferentes métodos de modelagem de tópicos para avaliação automática de grandes corpos de texto. O artigo de [Jankowski-Lorek and Zieliński 2015] avalia a controvérsia de um dado texto através de uma comparação do mesmo com a enciclopédia colaborativa *wikipedia* [6]. Para isso o autor, utilizou-se de uma métrica de similaridade de cosseno entre os vetores de *Time frequency - inverse document frequence (TF-IDF)* que representam os documentos, sendo possível aplicar este medida para tópicos semelhantes.

O trabalho de [Cheng et al. 2014] propõe uma nova abordagem para modelagem de tópicos e textos curtos, BIT *(Biterm topic model)*. Este algoritmo descobre tópicos modelando estes diretamente na geração de padrões de palavras que co-ocorrem nos textos, apresentando em pequenos textos tópicos mais coerentes. O artigo de [Zhao et al. 2011] é utilizada a MT para comparação de mídias tradicionais (jornais, revistas) com postagens do *twitter*, sendo utilizado no trabalho o algoritmo *Twitter-Latent Dirichlet Allocation (Twitter-LDA)* com a justificativa de que o algoritmo original, *Latent Dirichlet Allocation* (LDA), não se adapta tão bem ao tamanho diminuto do corpo de um *tweet*. O trabalho

[6] pt.wikipedia.org

[Luiz et al. 2018] se assemelha bastante com a proposta deste trabalho, sendo este um *framework* para análise de *reviews* de aplicativos de celular.

3. METODOLOGIA

O presente trabalho constitui da avaliação automática de conteúdo de sites de reclamação online, bases governamentais e redes sociais, utilizando técnicas de mineração de dados. Devido a imensa quantidade de dados disponíveis nesses sites e bases, decidiu-se por restringir o número de plataformas de reclamação e a rede social utilizada, sendo utilizado o ReclameAqui, que hoje é uma das maiores páginas do Brasil de reclamações online, como também a rede social Twitter, por sua facilidade de compartilhar e receber informações.

Também foram restringidas as empresas a serem estudadas, já que há uma grande variabilidade de empresas presentes nestas base, sendo inviável a análise de todas para este trabalho. Desta forma, foram selecionadas como alvo as quatro maiores companhias de telefonia do Brasil, sendo elas Oi, Tim, Vivo e Claro. A motivação para escolhermos tais empresas se dá pela grande quantidade de usuários que as mesmas possuem, sendo mais de 235 milhões de linhas móveis atualmente [7]; a variedade de serviços oferecidos, como telefonia, internet móvel, TV a cabo, entre outros; e, além disso, o número de reclamações presentes nessas bases contra companhias do setor de telefonia, sendo registradas mais 28 mil reclamações fundamentadas pelo PROCON somente no ano de 2016. Logo, a metodologia proposta para esta análise se divide em: coleta de dados, pré processamento, análise de sentimentos e modelagem de tópicos.

A primeira parte do trabalho se dá pela coleta de dados para posterior análise. Foram coletados dados da plataforma de reclamações online ReclameAqui, dados governamentais de reclamações fundamentadas [8] e Twitter [9]. Os dados governamentais estavam presentes em formato semi-estruturado, havendo dados de 2009 a 2016 em formato *csv*, podendo ressaltar que estes são abertos disponíveis em [10]. Um *crawler* foi implementado para realizar a busca e armazenamento das reclamações relacionadas com companhias telefônicas do ReclameAqui; já para a coleta dos *tweets* relacionados a empresas de buscas, foi utilizada a biblioteca *tweepy* [11]. A busca foi realizada nos *twitters* oficiais das companhias, assim como no *twitter* oficial da Anatel.

Para a base de dados do PROCON, foi feita uma caracterização da base. A caracterização de uma base é definida como em uma descrição dos atributos de uma base, identificando padrões básicos, sem a necessidade de aplicação de um algoritmo específico. Optou-se por este tipo de abordagem nesta base, visto que as reclamações presentes nela se mostravam previamente classificadas, sendo desta forma desnecessária a aplicação de um algoritmo de *text mining* para análise.

A segunda parte de nossa metodologia compreende o pré processamento dos dados. Esta etapa se mostrou vital para a avaliação dos dados, e que estes fossem obtidos em um tempo hábil, já que foi reduzida a dimensionalidade das frases a se tratar retirando algumas classes de palavras do texto. Para a fase de pré processamento, realizada em todas as bases, foi utilizada a biblioteca *NLTK* [12], que disponibiliza diversas funções para tratamento da base, como remoção de *stop words*, pontuação e lematização das palavras presentes. Por se mostrar um modo mais informal para compartilhamento de texto, os *tweets* apresentavam diversos *emotions* e caracteres não alfa-numéricos, que foram retirados por afetarem de maneira negativa as próximas etapas.

A terceira etapa do trabalho corresponde a análise sentimental dos textos. A base de *tweets* pode

[7] www.anatel.gov.br/dados/destaque-1/283-brasiltem-236-2-milhoes-de-linhas-moveis-em-janeiro-de-2018
[8] dados.gov,br
[9] twitter.com
[10] http://dados.gov.br/dataset/cadastro-nacional-de-reclamacoes-fundamentadas-procons-sindec1
[11] tweepy.org
[12] nltk.org

apresentar vários tipos de texto que se relacionam com os mais diversos tópicos, desde propagandas, elogios, até nosso alvo, que são as reclamações. Considera-se que essas possuem um sentimento negativo agregado a elas, e assim, através da análise de sentimentos, é possível selecionar apenas os *tweets* que possuem uma "emoção"negativa, abrangendo a grande maioria das reclamações presentes no Twitter.

Para análise de sentimentos foi utilizado um método baseado em *lexicons*, utilizando a ferramenta *VADER (Valence-Aware Dictionary for sEntiment Reasoning)* [Gilbert 2014]. De maneira geral, o VADER realiza a junção de diversos léxicos de outras abordagens de análise de sentimentos, como *LIWC, ANEW, SentiWordNet*, entre outros. O VADER é atualmente considerado o estado da arte, sendo possível destacar que este retorna além do sentimento relacionado a um corpo de texto, a intensidade relacionada aquela emoção.

A quarta e última parte do trabalho se dá pela modelagem de tópicos. Para realização desta parte foi realizada a remoção temporária de palavras que possuem sentimento. O objetivo desta remoção é a redução de ruídos que possam ser gerados durante a identificação dos tópicos, visando obter uma resposta mais precisa daquilo que a reclamação realmente representa.

Para a modelagem de tópicos foi utilizado o algoritmo LDA, sendo este amplamente utilizado na literatura e bastante apropriado para descoberta de tópicos em corpos de texto. Este emprega uma técnica não supervisionada de aprendizado de máquina que identifica tópicos latentes de informação em vastas coleções de documentos [Hong and Davison 2010]. O algoritmo utiliza-se de uma abordagem de *Bag Of Words* (BOW), tratando cada documento como um vetor de palavras contadas. Cada documento é representado pela probabilidade de distribuição em um número de palavras.

4. RESULTADOS

4.1 Bases de dados

Em nossa avaliação experimental foram utilizadas três bases de dados coletadas com reclamações das quatro maiores operadoras de telefonia do Brasil. As bases são provenientes do PROCON, do *Twitter* e do ReclameAqui. Vale destacar que essas bases de dados se mostram bastante distintas, sendo representadas por tópicos de reclamações fundamentadas; textos curtos, que podem possuir reclamações ou não (exigindo assim um tratamento melhor da base); e por textos que possuem uma variação de tamanho (curto ou longo). A Tabela I apresenta a quantidade de instâncias de dados coletados em cada base de dados utilizada, assim como a quantidade de reclamações por empresa em cada base:

Base	Quantidade de instâncias	Claro	Oi	Tim	Vivo
Twitter	67.916	21,6%	21.87%	26,34%	30,31%
Reclame aqui	40.000	25%	25%	25%	25%
Procon	187.925	26,60%	47,06%	15,27%	11,05%

Table I. Quantidade de dados coletados por base utilizada

4.2 Avaliação de reclamações: PROCON

A fim de caracterizar e entender de maneira melhor a base do PROCON, foram separadas as reclamações únicas de todas as bases. Vale destacar que nesta base não foi necessária a aplicação da modelagem de tópicos, já que as reclamações já foram categorizadas previamente pelo PROCON. Desta forma, apenas com a caracterização da base, foi possível descobrir as maiores reivindicações contra cada uma das empresas. A partir disso foram identificados 100 tópicos distintos, considerando todas as bases do PROCON. Entretanto, a base não foi gerada com devidos cuidados, sendo encontrado diversas reclamações que não condizem com serviços prestados por operadoras de telefonia, como

por exemplo: "Água/Esgoto", "Óticas (Lentes/Óculos)", "Transporte Escolar", entre outros. Desta forma, se mostrou necessária a remoção dos diversos tópicos não associados a empresas de telefonia, sendo ao final identificados apenas 18 tópicos válidos ao contexto. Dos tópicos válidos, foi possível identificar que alguns tinham o mesmo significado no cenário de estudo, como por exemplo: "Telefone (Convencional, Celular, Interfone, Etc.)", "Telefone Celular", "Telefonia Fixa". A Tabela II apresenta os tópicos gerados.

Tópicos	Claro	Oi	Tim	Vivo
Banco de Dados (SPC-SERASA-ETC)	32.87 %	37.33 %	19.82 %	9.98 %
Cartão de Crédito	20.85 %	54.4 %	17.46 %	7.29 %
Empresa de cobrança	22.77 %	57.28 %	12.68 %	7.28 %
Financeira	33.8 %	42.7 %	13.0 %	10.5 %
Internet	38.94 %	37.97 %	9.76 %	13.34 %
Microcomputador / Produtos de Informática	38.92 %	15.95 %	21.69 %	23.44 %
Outros Contratos	29.36 %	43.96 %	15.7 %	10.98 %
Serviços Telefônicos Especiais (Disque 900, Erótico, Etc)	21.47 %	68.05 %	0.0 %	10.48 %
Telemarketing	58.65 %	28.85 %	0.0 %	12.5 %
TV por assinatura	63.08 %	34.1 %	0.2 %	2.62 %
Telefonia Geral (Fixa, Móvel)	22.71 %	49.23 %	16.78 %	11.28 %
Telefonia Comunitária (PABX, DDR, Etc.)	20.53 %	64.26 %	5.64 %	9.56 %

Table II. Acima estão descritos os 12 tópicos mais importantes da base de dados do PROCON e número de reclamações relacionada ao tópico em cada uma das empresas

Ao analisarmos os tópicos e a quantidade com que eles ocorrem (Tabela II), vemos que grande maioria das reclamações estão relacionadas com os tópicos "Telefonia Geral (Fixa, Móvel)"e "Internet". Entretanto, percebe-se também que os tópicos se mostram superficiais e com pouquíssimas informações, e que a descrição do problema se mostra resumida e não condizente com as categorias descritas. Tendo em vista esses aspectos levantados, pode-se concluir que essa base se mostra desprovida de informações substanciais que possam auxiliar empresas a resolver problemas, uma vez que ela apresenta as categorias de maior reclamação, mas não detalhes suficiente para tratá-los.

4.3 Avaliação de reclamações: ReclameAqui

Na base do ReclameAqui foi realizado o pré-processamento e utilizado o algoritmo LDA de modelagem de tópicos, visando a identificação de tópicos latentes que possam sumarizar, de maneira fidedigna, tudo que está presente nos dados. Vale ressaltar que as reclamações presentes nessa base se caracterizam pela grande variação do tamanho dos textos que a compõe, dispondo de passagens muito grandes, até sentenças com poucas palavras. Após a aplicação da MT, foram identificados os seguintes tópicos: (1) Internet; (2) Sinal; (3) Portabilidade; (4) Atendente; (5) Fatura; (6) Técnico; (7) Ligações; e (8) Pagamento. Através dos tópicos, pode-se observar que foi possível captar os principais problemas que ocorrem em companhias telefônicas, e de maneira oposta a base de dados vinda do PROCON, esta base pode oferecer informações substanciais que auxiliem empresas no tratamento de problemas específico.

4.4 Avaliação de reclamações: *Twitter*

A base com dados provenientes do *Twitter* foi criada por meio da análise de *tweets* direcionados as grandes operadoras de telefonia e a ANATEL. Como o *Twitter* é uma rede social, ele permite que seus usuários postem diversos tipos de textos. Tal liberdade, permite que o usuário expresse, de forma mais fiel, o seu real sentimento, contudo existem desvantagens como a não-padronização dos textos escritos, sendo utilizadas figuras de linguagens (gírias), abreviações, *emotions* e até mesmo palavras erradas gramaticalmente.

Todos estes motivos apresentados, e o fato de estarmos analisando apenas reclamações, interferem de forma negativa na modelagem dos tópicos de redes socais em geral. Desta forma, para um melhor resultado dos tópicos gerados em bases de redes sociais, foram feitos alguns tipos de "pré-processamentos

especiais", como a retirada de jargões utilizados no *Twitter*, tais como, a expressão "RT"*(retweet)*, expressões que começavam com "#"*(hashtag)*, entre outros. O nosso foco neste trabalho foi apenas a análise de reclamação, portanto considera-se que há um valor intrínseco de negatividade associado as reclamações postadas. Desse modo, a análise de sentimentos auxilia na recuperação apenas de *tweets* com um sentimento negativo associado, representando assim uma reclamação.

Para análise de sentimentos neste trabalho foi utilizada a ferramenta VADER, aplicada como um filtro na recuperação de textos com sentimentos negativos apenas. Vale destacar, que dos aproximadamente 68 mil *tweets* coletados, apenas cerca de 24 mil possuíam o sentimento negativo agregado a eles, isto é, em torno de 35% do total. Entretanto, devido ao pequeno tamanho dos corpos dos documentos presentes no *Twitter*, máximo de 280 caracteres, a modelagem de tópicos seria prejudicada pela pequena quantidade de documentos, caso fosse feita a análise por operadora. Em virtude destes fatos, foi aplicada a modelagem de tópicos sobre todos os documentos, sem distinção de operadora, para que pudéssemos aumentar a coerência dos tópicos gerados, e termos uma visão geral das reclamações feitas sobre todas as operadoras. Os tópicos relacionados com reclamações de operadoras sobre a base de dados foram: (1) Chip; (2) Fibra; (3) Técnico; (4) Internet; (5) Fatura; (6) Minutos; (7) Atendimentos; e (8) Ligação.

4.5 Discussão dos Resultados

Foi realizada uma comparação dos principais tópicos extraídos do ReclameAqui relacionados a cada uma das empresas de telefonia separadamente, utilizando a correlação de *Spearmam*. Conforme podemos observar na Tabela III, algumas empresas possuem as mesmas reclamações que outras, devido a grande correlação entre essas empresas. Este caso ocorre em 3 das grandes operadoras que possuem uma correlação maior de 0.3 entre elas, o que gera um respaldo ao consumidor caso ele queira mudar de operadora, mostrando que o mesmo não possui muitas opções, já que as adversidades que ocorrem em uma companhia também ocorrem em outra. O que diferencia é a intensidade em que essas adversidades ocorrem.

Operadoras	Claro	Oi	Tim	Vivo
Claro	1	-0.37	0.35	0.40
Oi	-0.37	1	0.16	0.28
Tim	0.35	0.16	1	0.35
Vivo	0.40	0.28	0.35	1

Table III. Correlação entre as bases

Com relação aos tópicos gerados, podemos observar que os tópicos extraídos automaticamente utilizando as bases de dados do ReclameAqui e do do Twitter são bem mais descritivos e relacionados a empresas de telefonia, quando comparamos com a taxonomia do PROCON. Por outro lado, apesar da semelhança entre os tópicos extraídos do ReclameAqui e do Twitter, o custo associado a base do Twitter é bem maior, tanto do ponto de vista computacional, na execução de diversas estratégias de pré-processamento, quanto do de recursos humanos, no tratamento de diversas exceções. Assim, a partir de bases de dados de aplicações de reclamação online, podemos extrair informações tão úteis quanto de redes sociais, mas utilizando menos recurso. Além disso, essas informação podem e devem ser utilizadas em complemento àquelas apresentadas pelo PROCON.

5. CONCLUSÕES E TRABALHOS FUTUROS

Nesse trabalho propomos uma metodologia para avaliação de comentários em aplicações de reclamação online. Combinando estratégias de preprocessamento de texto, análise de sentimento e modelagem de tópicos, o objetivo é extrair informações úteis, ricas em detalhes, que possam contribuir para empresas identificarem de forma mais consistente e rápida problemas em seus produtos e/ou serviços. Avaliamos

nossa metodologia considerando uma coleção de comentários relacionados às quatro maiores empresas de telefonia do Brasil (TIM, OI, CLARO e VIVO) extraídos automaticamente a partir da aplicação ReclameAqui. Comparamos os temas de reclamação identificados por nossa metodologia com as reclamações cadastradas no PROCON e pudemos constatar que os tópicos extraídos automaticamente pela metodologia a partir dos comentários dos usuários são mais consistentes e mais refinados que os cadastrados no PROCON. Além disso, aplicamos também nossa metodologia em uma coleção de *tweets* relacionados as mesmas empresas e observamos que os tópicos identificados eram muito próximos daqueles relacionados a coleção extraída do ReclameAqui, porém foram necessários mais etapas de pré-processamento de texto um vez que coleções oriundas de redes sociais são carregadas de ruídos. Nossa conclusão final é que os comentários de aplicações de reclamação online se apresentam como a melhor alternativa para se identificar e extrair problemas relacionados a produtos e/ou serviços.

Como trabalho futuro, nosso objetivo é a construção de um *framework* que possa fornecer a empresas informações mais detalhadas sobre suas reclamações, criando um modo atrativo de visualização que permita a geração intuitiva dos dados coletados. Também pretendemos empregar diferentes algoritmos de modelagem de tópicos, como: *Non-Negative Matrix Factorization*, *Singular Value Decomposition*, BTM, entre outros. Espera-se aplicar exclusivamente uma técnica de modelagem de tópicos que consiga ter uma alta coerência em pequenos corpos de texto. Por último, pretendemos desenvolver um *Chatbot*, que após treinado, possa responder automaticamente a reclamações, dando ao consumidor um respaldo em relação a sua reivindicação.

Agradecimentos

Esse trabalho foi parcialmente financiado por CNPq, CAPES, FINEP, Fapemig, MASWeb e INWEB.

REFERENCES

Almeida, T. G., Souza, B. A., Menezes, A. A., Figueiredo, C., and Nakamura, E. F. Sentiment analysis of portuguese comments from foursquare. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*. ACM, pp. 355–358, 2016.

Auden, W. H. *The Complete Works of W. H. Auden: Prose*. Vol. 2. Princeton University Press, 2002.

Cheng, X., Yan, X., Lan, Y., and Guo, J. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering* 26 (12): 2928–2941, 2014.

Gilbert, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

Hong, L. and Davison, B. D. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*. ACM, pp. 80–88, 2010.

Jankowski-Lorek, M. and Zieliński, K. Document controversy classification based on the wikipedia category structure. *Computer Science* vol. 16, 2015.

Luiz, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M. A., and Rocha, L. A feature-oriented sentiment rating for mobile app reviews. In *Proceedings of the 2018 World Wide Web Conference*. WWW '18. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1909–1918, 2018.

Pak, A. and Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*. Vol. 10, 2010.

Rocha, L., Mourão, F., Silveira, T., Chaves, R., Sa, G., Teixeira, F., Vieira, R., and Ferreira, R. Saci: Sentiment analysis by collective inspection on social media content. *Web Semantics: Science, Services and Agents on the World Wide Web* vol. 34, pp. 27–39, 2015.

Sá, G., Silveira, T., Chaves, R., Teixeira, F., Mourão, F., and Rocha, L. Legi: Context-aware lexicon consolidation by graph inspection. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. ACM, pp. 302–307, 2014.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*. Springer, pp. 338–349, 2011.