Clusterização de Reclamações da Plataforma Cidadão Reclame Aqui

Arthur Um
No USP: 9778898
arthur.um@usp.br

Augusto Calado Bueno
No USP: 9779134
augusto.bueno@usp.br

Fernando Chiu Hsieh
No USP: 9436743
fernando.hsieh@usp.br

João Trevisan Martins
No USP: 9778599
joao.trevisan.martins@usp.br

Lucas Shie Lai
No USP: 9778710
lucas.lai@usp.br

I. Introdução

Atualmente, com o advento das tecnologias de informação houve a criação de canais de comunicação acessíveis para grande parte da população. Uma das finalidades dadas a esses canais foi o desenvolvimento de plataformas de comunicação entre cidadãos e governo.

O Procon-SP é um portal com a finalidade de servir como canal de reclamações em que os cidadãos podem denunciar irregularidades provenientes de prestações de serviços ou produtos para o governo do estado de São Paulo, criando assim uma extensa base de dados de textos relacionados a uma ampla diversidade de áreas e assuntos. Essas reclamações podem trazer valiosas informações para a gestão do estado à medida que explicitam falhas que necessitam de políticas públicas para serem resolvidas ou, pelo menos, suavizadas.

Como essa plataforma atende o estado mais populoso do Brasil e, com a ampla disseminação da tecnologia, cada vez mais pessoas têm acesso a ela, dar vazão a todas essas reclamações separadamente torna-se um trabalho cada vez mais inviável. Pensando nesse cenário, a necessidade de criação de ferramentas que possibilitassem o tratamento desses textos por agrupamentos feitos de forma automática e eficiente surgiu. Dessa maneira, políticas públicas mais abrangentes e, por consequência, possivelmente mais eficazes poderiam ser tomadas após a análise dos dados processados.

Uma vez que a base de dados composta por todas as reclamações da plataforma governamental do Procon-SP é fechada, outras alternativas tiveram de ser consideradas para a criação do modelo a ser desenvolvido. Após uma análise realizada sobre algumas plataformas de finalidades análogas, chegou-se a decisão do uso de textos de reclamações retirados da plataforma Cidadão Reclame Aqui, isso devido ao seu foco ser fortemente parecido ao escopo do Procon-SP.

Para a extração dos textos da plataforma Cidadão Reclame Aqui foi necessário o uso de um *Crawler*, basicamente uma ferramenta que percorria página por página da plataforma retirando os textos das reclamações e outros dados como categoria, data, título e local. Por fim, esses dados foram salvos em arquivos separadamente para uso posterior.

Portanto, o objetivo final do estudo pode ser resumido em criar métodos automáticos baseados em Processamento de Linguagem Natural de agrupamento de reclamações do Cidadão Reclame Aqui, passíveis de análise qualitativa para possível aproveitamento em ferramentas governamentais de forma a auxiliar em tomadas de decisão dos gestores.

II. Conjunto de Dados

O conjunto textual extraído para o estudo é composto de 94557 documentos, com cada um composto de uma única reclamação e outros dados referentes a ela (data, local, título e categoria).

Essas reclamações são referentes a datas que variam entre abril de 2013 e setembro de 2018, e pertencem a 83 categorias no total.

Como 83 categorias é um número muito elevado para a tarefa de agrupamento, gerando assim uma complexidade superior à aceitável, foi decidido por gerar manualmente "super-categorias", basicamente novas categorias compostas por categorias anteriores que compartilham o mesmo contexto. Um exemplo seria textos que pertenciam a categorias como Banda Larga Fixa, TV por assinatura, Telefonia Fixa e Telefonia Celular, as quais passaram a ser tratadas como parte da nova super-categoria chamada de Telecomunicações.

O nome e composição de cada uma das categorias geradas manualmente segue abaixo:

1) Atendimento (23646 reclamações):
   - Atendimento
   - Atendimento da ANA
   - Atendimento da ANAC
   - Atendimento da ANP
   - Atendimento da ANS
   - Atendimento da ANATEL
   - Atendimento da ANCINE
   - Atendimento da ANVISA
   - Atendimento de Órgãos
   - Comunicação com o usuário
   - Denúncia
   - Denúncia de Estabelecimentos
   - Serviços
   - Serviços aos Clientes
2) Telecomunicações (10583 reclamações):
   - Banda Larga Fixa
   - TV por assinatura
   - Telefonia Fixa
   - Telefonia Celular

3) Saúde (4313 reclamações)
   - Ambulatórios
   - Hospitais
   - Postos de Saúde
   - Medicamentos, Produtos e Alimentos
   - Prestação de Serviços com risco à Saúde
   - Zoonoses
4) Educação Pública (2222 reclamações)
   - Escola Estadual
   - Escolas Municipais
   - Faculdades Estaduais
   - Creches
5) Infraestrutura (11442 reclamações)
   - Aeroporto
   - Buracos
   - Estradas
   - Expansão de Faixas e Corredores
   - Iluminação Pública
   - Obras Públicas
6) Meio Ambiente (9487 reclamações)
   - Árvores
   - Limpeza de Terrenos
   - Lixo e Poluição
   - Meio Ambiente
   - Meios
   - Poluição Sonora
7) Transporte (13215 reclamações)
   - Acessibilidade
   - Acessibilidade - Geral
   - Acessibilidade - Transporte
   - Bilhetagem
   - Concessões Ferroviárias
   - Terminais, estações e Pontos
   - Transporte
   - Transporte Público
   - Transporte Rodoviários de Cargas
   - Transporte Rodoviários de Passageiros
   - Trânsito
   - Voos Domésticos
   - Voos Internacionais
8) Governamental (7117 reclamações)
   - Cidades
   - Estados
   - Estadual
   - União
   - Cultura, Esporte e Lazer
   - Segurança
   - Segurança - Estadual
   - Segurança - Geral
   - Contratos e Regulamentos
   - Empregados
   - Oportunidades de Emprego
   - Trabalho e Emprego
   - Cobrança
   - Impostos Estaduais
   - Impostos Federais
   - Impostos e Taxas
   - Mensalidades e Reajustes
9) Outros (12475 reclamações)
   - Outros
   - Convívio Público
   - Certificação
   - Exibição
   - Financeiro
   - Fiscalização
   - Institucional
   - Geral
   - Cadastro
   - Revenda
   - Programas

A super-categoria Outros é composta por todas as categorias anteriores que não se encaixavam em nenhuma outra super-categoria. Como isso poderia ser danoso para a tarefa de clusterização, optou-se por retirar todos os textos que a compunham, resultando assim em um número final de 82082 textos.

Tendo em vista o tamanho grande do córpus total, optou-se por restringir a análise somente para o ano de 2017, contando com um recorte de 13601 documentos.

III. Trabalhos relacionados

A. Trabalhos relacionados ao tema

1) Topic Classification And Clustering on Indonesian Complaint Tweets for Bandung Government using Supervised And Unsupervised Learning [1]: Assim como a proposta de analisar as reclamações registradas no PROCON utilizando técnicas de processamento de linguagem natural, o artigo em questão foca, também, no emprego de técnicas de Inteligência Artificial na análise e monitoramento das reclamações registradas no portal da cidade de Bandung, Indonésia para auxiliar no atendimento e resolução realizado pelo governo.

O sistema desenvolvido no artigo utiliza técnicas de aprendizado supervisionado para classificação de reclamações e aprendizado não-supervisionado (método utilizado pelo grupo para analisar as reclamações) para agrupar as reclamações baseando-se na informação contida em uma reclamação.

No aprendizado supervisionado, os pesquisadores empregaram o uso dos algoritmos Classificador Naive Bayes, Random Forest e Otimização mínima sequencial. Como medida de avaliação, utilizou-se F1-Score. Já no aprendizado não-supervizionado, os algoritmos utilizados foram *Document Pivot Tecnique E Exemplar Based Topic* usando TF-IDF como forma de representação dos dados, e para o cálculo da similaridade entre os texto utilizou-se a similaridade cosseno.

De acordo com os autores, os resultados alcançados atingiram bons valores. Concluíram que, para o caso de clusterização, o algoritmo *Exemplar Based Topic* obteve os melhores resultados; já para a tarefa de classificação single-label, o melhor algoritmo foi o Otimização Mínima sequencial e o

melhor classificador multilabel foi o *Random Forest*, com *F1 Score* de aproximadamente 0.98.

2) From Social Media to Public Health Surveillance: Word Embedding based Clustering Method for Twitter Classification [2]: A proposta discutida e apresentada no artigo é a utilização de *tweets* para determinar e prever a saúde da população. Uma das abordagens utilizadas no artigo para atingir o objetivo de determinar a saúde populacional é a aplicação de técnicas de processamento de linguagem natural, algumas das quais se relacionam com as utilizadas no projeto como, por exemplo, a utilização do *Word Embedding* para criar as representações das palavras dos documentos.

Os métodos para o processamento dos *tweets* foram divididos em três etapas: a primeira consiste no pré-processamento dos *tweets*, já que há diversas construções linguísticas que podem provocar ruídos nos resultados (*hashtag*, palavras abreviadas etc). Na etapa dois, realiza-se o processo de clusterização utilizando o algoritmo *Chinese Restaurant Process* (algoritmo utilizado na resolução do presente trabalho com uma abordagem voltada para o pós-processamento). A terceira etapa consiste em identificar a quais clusters cada documento pertence, para isso eles usam *similarity measure*.

O objetivo dos pesquisadores era determinar uma maneira diferente e mais eficiente para capturar a semântica nos *tweets* para que, dessa forma, fosse possível determinar eventuais problemas de saúde na população. Para tal, basearam-se no método de *word embedding* que, através da representação das palavras em vetores, permite capturar a semântica da palavra. Os resultados obtidos, de acordo com a opinião dos pesquisadores e, também, com as medidas de avaliação (como *F1 Score* igual a 84,7% para o *threshold* em 0.5) foram de grande qualidade. Concluíram que o método estudado é bem maleável para diversas áreas. Por ser um método não-supervisionado, eles ressaltam que um dos benefícios desse tipo de abordagem é a dispensa de classificação e treinamento do algoritmo.

B. Trabalhos relacionados às técnicas

1) Word Embedding: Essa técnica baseia-se em uma representação de palavras espalhadas em um espaço n-dimensional (nesse caso, com n igual a trezentos). Essa representação é gerada a partir de uma rede neural rasa. Palavras que aparecem em contextos parecidos durante o treinamento da rede aproximam-se também na representação. Com isso, vetores agrupam-se gerando espaços que compartilham significados semânticos semelhantes em relação aos termos que os compõem [3] [1].

[1]Durante o nosso trabalho intitulamos essa representação de word2vec por conta do nome do algoritmo utilizado para treinar os embeddings.

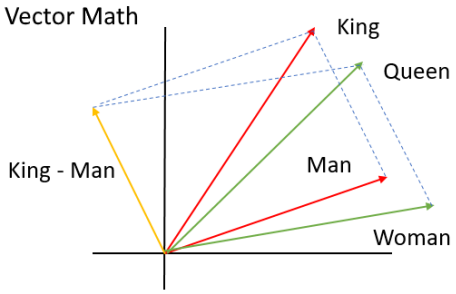

Figura 1. Exemplo de representação de palavras por vetores assim como é feita com o *Word Embbedings*

2) Latent Dirichlet Allocation: Latent Dirichlet Allocation (LDA) [4] é um algoritmo probabilístico utilizado para descobrir os tópicos contidos em conjuntos de palavras. No contexto deste artigo, o LDA representa cada documento como uma mistura de tópicos que produzem as palavras utilizadas, sendo que cada palavra possui uma probabilidade diferente de ser produzida dado o tópico. Desta forma, o algoritmo assume que palavras que pertencem ao mesmo contexto (documento), ou a contextos do mesmo gênero, pertencem aos mesmos tópicos também.

Dado um córpus e um número de tópicos T, uma das formas que o algoritmo atribui cada uma das palavras e documentos aos tópicos é através dos seguintes passos:

1) Ao passar por cada documento, atribuir cada palavra a um dos T tópicos de forma aleatória.
2) Novamente passando por cada documento d, iterar por cada palavra p do documento e calcular as seguintes probabilidades: P(tópico t | documento d) e P(palavra p | tópico t) para todos os tópicos, e atribuir a palavra a um novo tópico, de acordo com a multiplicação das probabilidades calculadas.
3) Repetindo o passo 2 várias vezes, até que tópicos mais bem definidos são gerados.

Como resultado, cada tópico gerado pelo algoritmo será composto por uma distribuição de probabilidade sobre todas as palavras recorrentes do córpus. Ao final de todas as iterações do algoritmo, estará pressuposto que os tópicos serão compostos por palavras semanticamente relacionadas entre si, e que fornecerão proporções maiores às palavras mais relevantes a si mesmos. Por exemplo, um tópico contendo duas palavras principais “cachorro-quente” e “comida”, tal que a palavra “comida” é a mais relevante, poderia ser representado por: 70% “comida”, 20% “cachorro-quente” e 10% outros.

Diferentemente dos clusters gerados pelo K-means ou algoritmo processo do restaurante chinês, os tópicos gerados pelo LDA são representados já pelas palavras, e não pelos documentos. Dessa maneira, o LDA realiza um soft-clustering, que computa os valores semânticos de grupos de palavras no contexto do corpus inteiro (utilizando os contextos dos documentos). Desta forma, não há a limitação contida nos clusters de documentos em que a representação de um dado cluster é dada apenas pelas palavras que compõem os documentos contidos no mesmo.

Levando em consideração que o córpus obtido para a execução deste trabalho é composto por vários temas diversificados de tal forma que a mistura de palavras-chaves em documentos é frequente, esta característica do algoritmo LDA poderá propor representações definidas em termos de proporções de tópicos computados, o que permite avaliar os documentos em uma perspectiva mais ampla, considerando-os em relação aos tópicos e de maneira mais global.

IV. Metodologia

Os métodos automáticos de agrupamento de reclamações pretendidos podem ser divididos em quatro etapas: o pré-processamento das reclamações; o agrupamento não supervisionado das reclamações; a avaliação da qualidade dos agrupamentos; e a extração de informações de cada agrupamento, semelhantemente ao trabalho realizado em [1]. A explicação de cada etapa é sucintamente mostrada a seguir:

1) Pré-processamento das reclamações: Nesta etapa, as reclamações passam por um processo de limpeza de dados, realizando-se a remoção de possíveis inconsistências dos textos (tais como erros tipográficos) e *stopwords* [2]. Após isso os textos são transformados em alguma representação numérica dos mesmos, as quais podem servir de entrada para os métodos de agrupamento não supervisionado. As representações numericas podem se basear em contagens de palavras (*Term Frequency Inverse Document Frequency* -TFIDF[3]- por exemplo) ou em modelos de representação distribuídas de palavras (*word embeddings*) explicado na seção anterior.
2) Agrupamento das reclamações: Nesta etapa, métodos de agrupamento não supervisionado são aplicados nas representações numéricas geradas anteriormente, obtendo-se agrupamentos dessas reclamações. Exemplos de algoritmos de agrupamento são: *K-means* e Processo do Restaurante Chinês [5].
3) Avaliação da qualidade dos agrupamentos: Nesta etapa, são empregadas métricas quantitativas para avaliar a qualidade dos agrupamentos. Exemplos de métricas quantitativas são índices internos que avaliam o grau de coesão e separação dos clusters (por exemplo o índice silhueta [6] ou índice Dunn).
4) Extração de informações dos agrupamentos: Nesta etapa, métodos de extração de informação são empregados para avaliar o conteúdo dos agrupamentos. Podem ser empregadas uma inspeção manual, análise de palavras mais frequentes ou reconhecimento de entidades e relações.

[2]Stopwords são palavras que são encontradas em grande frequência nos textos mas que não proporcionam um ganho de informação para os modelos. Pelo contrário, elas atrapalham servindo de ruídos. Exemplos de stopwords são artigos e preposições.

[3]TFIDF é uma maneira de representação de palavras que realiza a contagem de termos em um documento ao mesmo tempo que atribui uma importância para cada um desses termos baseado na frequência de cada um deles no córpus.

V. Descrição dos métodos propostos

A fim de determinar qual método de agrupamento automático é melhor indicado para tratar as reclamações do Reclame Aqui, foi feita uma análise exploratória, testando-se diferentes formas de representação de textos, algoritmos de agrupamento e métodos de extração de informação.

Foram utilizadas duas formas de representação dos textos: uma representação TFIDF dos textos considerando somente as 1000 palavras mais frequentes e uma representação que utiliza os *word embeddings*, tomando todas as palavras de um documento e representando-o pela média dos vetores dessas palavras. Foram usados os *embbedings* de trezentas dimensões treinados utilizando a técnica *skip-gram* retirados do NILC [7]. Além disso, removidas as *stopwords* dos textos segundo a lista de *stopwords* para o português do NLTK (Natural Language ToolKit) [8].

Para o agrupamento foram considerados dois algoritmos de aprendizado não supervisionado: *K-means* e LDA. Para avaliar quantitativamente a qualidade dos clusters foi utilizado o índice silhueta e uma contagem da proporção de cada categoria dentro desse cluster. Para a extração de informação foi utilizada uma análise de palavras mais frequentes além de uma variação do processo do restaurante chinês que retorna agrupamentos menores de palavras dentro de cada cluster (esta variação é baseada no trabalho realizado em citeHealth), intitulado aqui por nós *WordChineseRestaurant*.

Os testes foram realizados variando o número de clusters no intervalo de 5 a 11 clusters. Intuitivamente foram escolhidos esses números a fim de chegar em resultados que fossem próximos à quantidade de super-categorias do córpus e que não fossem excessivamente altos por conta da dificuldade de análise que isso proporcionaria.

A. Descrição do WordChineseRestaurant

Esse método de agrupamento de palavras é baseado no trabalho realizado por [2] e fundamenta-se no algoritmo do Processo do Restaurante Chinês e em uma métrica de similaridade entre palavras (computada com o uso de *word embeddings*).

No primeiro momento, são concatenados todos os textos relativos a um cluster, logo mais gerando uma lista de todas as palavras de todos os textos daquele cluster. A seguir a primeira palavra da lista é associada a primeira mesa (subcluster). A partir daí, cada palavra subsequente da lista ou é atribuída a uma nova mesa segundo uma probabilidade p que decresce com o avanço do algoritmo, ou a uma mesa já estabelecida de acordo com maior similaridade. A similaridade entre as palavras foi calculada pela similaridade cosseno entre os vetores das mesmas (provenientes dos *word embeddings*). O pseudocódigo do algoritmo é apresentado abaixo:

Algoritmo 1: WORDCHINESERESTAURANT

$n = 1$
$p = 1/(1+n)$
Associa a primeira palavra à primeira mesa
loop até o fim da lista
1) sorteie um r entre $(0, 1)$
2) se $r < p$
   a) gere novo cluster
   b) $n = n + 1$
   c) atualize p
   d) adiciona a palavra à nova mesa
3) **else**
   a) compute a similaridade entre o vetor da palavra e a média dos vetores de cada mesa
   b) adicione a palavra àquela mesa que maximize essa similaridade.

retorna as mesas

B. Descrição do Word Cloud

Para que fosse possível interpretar os agrupamentos que viriam a ser criados posteriormente foi optado por apresentá-los por meio de nuvens de palavras.

Essas nuvens são apenas uma representação visual das palavras mais frequentes de cada um dos agrupamentos. Uma outra particularidade que optou-se foi apresentar as palavras mais frequentes em um tamanho proporcionalmente maior. Dessa forma, as palavras mais fáceis de se observar nas imagens são as que mais aparecem nos documentos ou tópicos clusterizados.

VI. ANÁLISE DOS RESULTADOS

Os resultados obtidos foram analisados sob a perspectiva de tentar encontrar o modelo que proporcionasse melhores exames qualitativos. Assim como será mostrado a seguir, isso fez com que fossem optados deliberadamente por investir em modelos que não necessariamente possuíssem os melhores índices internos.

A. Resultados do K-means

1) TFIDF: A seguir são apresentamos os resultados obtidos após a execução do algoritmo K-means, considerando a representação *TFIDF*. Os resultados dos índices silhueta variando o número de cluster são mostrados a seguir.

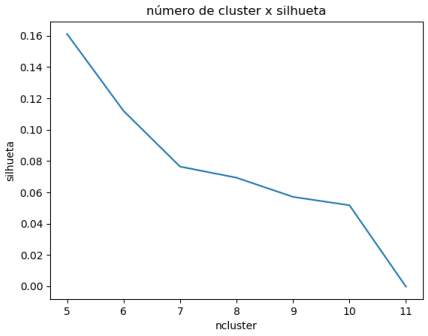

Figura 2. Gráfico de número de clusters x silhueta

Pode-se notar que conforme aumentou-se o número de clusters houve uma queda nos valores de silhueta. Isso colabora com a hipótese de que os dados estão homogeneamente distriuídos no espaço.

Quanto menor for o número de clusters, obtem-se agrupamentos mais densos com um grande número de indivíduos. Complementarmente, quanto maior o número de cluster menos indivíduos por cluster haverá. Essa proposição ajuda a pensar na hipótese de que um menor número de clusters proporciona uma generalização de conhecimento, pois cada agrupamento é utilizado para representar uma ampla gama de indivíduos. Acredita-se que idealmente deseja-se obter agrupamentos pequenos de indivíduos que apresentem um perfil particular e específico, em detrimento de uma grande massa que proporcione um conhecimento genérico.

A partir daí segue-se com a heurística de privilegiar números de cluster mais altos, contrariando a tendência de estimar números de cluster baixo segundo o índice silhueta.

Para colaborar com essa heurística, o grupo de fato considerou em geral os resultados qualitativos de agrupamentos com $ncluster$ altos melhores do que resultados de agrupamentos com $ncluster$ baixos.

Abaixo são mostrados algumas nuvens de palavras resultantes do agrupamento utilizando $ncluster = 11$.

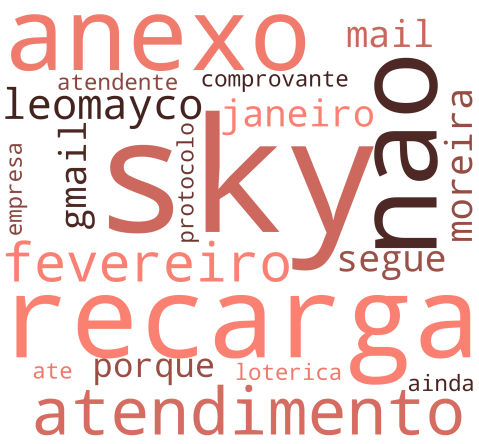

Figura 3. Cluster(7) do outlier

Figura 4. Cluster(8) da Telecomunicação

Figura 5. Cluster(11) do Atendimento

A análise das palavras mais frequentes desses agrupamentos permite concluir que o algoritmo conseguiu separar razoavelmente dois temas: um de telecomunicações (com 93.24% de telecomunicações) e outro aparentemente relacionado a atendimento.

Tabela I
PROPORÇÃO DAS CATEGORIAS DE COMPOSIÇÃO DOS CLUSTER REFERENTES AO EXPERIMENTO USANDO TFIDF

Categoria	Cluster 7	Cluster 8	Cluster 11
atendimento	0.00%	5.70%	5.55%
telecomunicações	100.00%	93.14%	93.22%
saúde	0.00%	0.06%	0.15%
educação pública	0.00%	0.00%	0.15%
infraestrutura	0.00%	0.30%	0.46%
meio ambiente	0.00%	0.18%	0.15%
transporte	0.00%	0.30%	0.00%
governamental	0.00%	0.30%	0.31%
total	1	1648	649

Vale observar que o algoritmo do kmeans em junção com o TFIDF estava muito sensível a *outliers*, por vezes agrupando somente um indivíduo. A figura ilustra esse fato, pois ela se refere a um cluster de um só indivíduo, o que explica a evidência de palavras pouco relacionadas.

2) Word2Vec: Os resultados a seguir são referentes à execução do K-means utilizando a representação Word2vec para o número de cluster igual a 10.

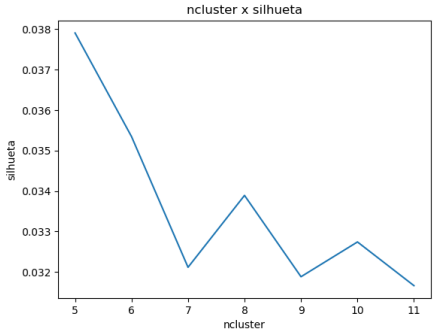

Figura 6. Gráfico número de cluster x silhueta

Figura 7. Cluster(8) da Saúde

Figura 8. Cluster(2) da Infraestrutura

Figura 9. Cluster(1) do Telecomunicação

Figura 10. Cluster(3) do Transporte

As quatro imagens anteriores relacionam de forma bem clara com alguma das categorias estabelecidas pelo grupo anteriormente. A nuvem de palavras de número seis se relaciona com o tema Saúde por apresentar palavras relacionadas à área como médico, consulta, postos e saúde. A grande maioria dos documentos que compõem o cluster em questão fazem parte das categorias atendimento (35,79%) e saúde (34,78%) e na nuvem de palavras nota-se, como dito anteriormente, encontrar que pertença a ambos os cluster . A sétima nuvem de palavras associa-se ao tema da infraestrutura, pois contém palavras como iluminação, lâmpada, queimada e apagada que juntas dão a idéia de queixas referentes a postes com lâmpadas queimadas ou apagadas e solicitações de trocas das mesmas. As categorias dos documentos que compuseram o cluster, em sua grande maioria estão categorizados como infraestrutura (83,97%).

O tema das telecomunicações é representado pela oitava nuvem de palavras, cujo principais palavras fazem alusão a serviços de internet, telefonia e empresas provedoras de serviço como a Vivo e NET. Dado os sentido semântico das outras palavras que compõem o cluster, nota-se a presença da ideia de problemas referentes a planos de celulares e internet. O cluster é composto por 80,43% dos documentos pertencentes à categoria telecomunicações.

Já a nona nuvem de palavras está altamente relacionada com o tema Transporte, pois sua composição conta com palavras chaves como ônibus, veículo, trânsito, motorista e transporte. A principal categoria no qual os documentos referentes a esse cluster pertencem é a de transporte que corresponde a 68.94% dos documentos.

Em relação ao índice silhueta, apesar do maior valor ser relacionado ao número de cluster igual a cinco, na prática, após análise qualitativas dos agrupamentos, interpreta-se que o índice silhueta não conseguiu ser um bom avaliador levando em conta um comparativo entre análises qualitativas (nuvem de palavras) e quantitativa (índice silhueta), em que nota-se uma falta de correspondência entre o melhor número de cluster indicado pelo silhueta como dito anteriormente.

Tabela II
PROPORÇÃO DAS CATEGORIAS DE COMPOSIÇÃO DOS CLUSTER REFERENTES AO EXPERIMENTO USANDO WORD2VEC

Categoria	Cluster 1	Cluster 2	Cluster 3	Cluster 8
atendimento	16.75%	3.59%	11.08%	35.79%
telecomunicações	80.43%	0.21%	0.31%	1.94%
saúde	0.26%	0.21%	0.10%	34.78%
educação pública	0.05%	0.42%	0.10%	12.81%
infraestrutura	0.72%	83.97%	6.42%	2.48%
meio ambiente	0.05%	6.75%	3.31%	3.42%
transporte	1.43%	0.42%	68.94%	3.80%
governamental	0.31%	4.43%	9.73%	4.97%
total	1952	474	966	1288

Realizando uma análise entre as nuvens de palavras do word2vec e do TFIDF, nota-se um grande aumento da similaridade entre os textos de cada agrupamento e da qualidade semântica geral dos clusters. Tais resultados podem ser atribuídos ao fato de a similaridade entre vetores do word2vec corresponder de certo modo à similaridade semântica real entre as palavras [3].

B. Resultados do WordChineseRestaurant

A ideia do grupo com a aplicação do WordChineseRestaurant era conseguir analisar os resultados dos agrupamentos com uma granularidade menor, podendo identificar subtópicos de reclamação dentro de cada agrupamento. Com isso, almejávamos conseguir separar tópicos menos presentes que poderiam ter sido ocultados com a análise de palavras mais frequentes.

Contudo, os resultados obtidos pelo grupo não conseguiram alcançar tal objetivo. Tendo em vista o tamanho do córpus, a nossa execução do algoritmo gerava muitos subclusters de saída (em média 60) o que tornava as análises qualitativas impraticáveis. Para tentar contornar isso, limitamos o número máximo de mesas para 15. Mesmo assim, os sub-agrupamentos resultantes não apresentavam qualidade de agrupamento semântico significativo.

Vale a pena notar que, apesar de não conseguirmos identificar subtópicos de reclamações, o algoritmo conseguiu agrupar palavras com conteúdo sintático semelhante mesmo com muitos ruídos. Como por exemplo, os agrupamentos abaixo:

1) Grupo de Nomes:
   - Lopez
   - Wesley
   - Barbosa
   - Macedo
2) Grupo de Adjetivos:
   - contestado
   - surpreendido
   - reclamado
   - proposto

Figura 11. Tópico de Telecomunicação

C. Resultados do Latent Dirichlet Allocation

No caso deste trabalho, os textos foram pré-processados através do algoritmo Term Frequency-Inverse Document Frequency. Portanto, as contagens geradas pelos tópicos do LDA serão pseudo-contagens que levam em consideração o inverso das frequências das palavras no córpus inteiro. Para incluir apenas as palavras mais relevantes do córpus, foi decidido criar a matriz de tf-idf com as 1000 *features* (ou palavras) mais frequentes. Consequentemente, os vetores dos tópicos gerado pelo LDA terão probabilidades para 1000 palavras.

Não obstante, a clusterização através dos tópicos ainda é possível, e acrescenta à análise uma outra alternativa viável. A sugestão de clusterização implementada pelo grupo é a seguinte. Cada documento pode ser representado por um vetor tf-idf, e cada vetor gerado pelo LDA representa um tópico. A associação de documentos a tópicos (consequentemente, clusters) pode ser realizada a partir da diferença entre esses vetores. Uma das maneiras de realizar tal procedimento é descrito nos seguintes passos:

1) Normalizar cada vetor da matriz de tf-idf em acordo com a norma l1.
2) Normalizar cada vetor dos tópicos em acordo com a norma l1.
3) Para cada vetor da matriz de tf-idf:
   - Calcular a sua distância Manhattan (também em acordo com a norma l1) com todos os tópicos.
   - Associá-lo ao tópico cujo vetor está mais próximo ao seu próprio de acordo com as distâncias calculadas no passo anterior.

Como os vetores possuem 1000 colunas (features), os cálculos das distâncias poderiam ser altamente influenciados por apenas alguns dos outliers em algoritmos cuja norma é l2 ou acima. Por esta razão, a distância Manhattan foi escolhida como exemplar neste caso.

Figura 12. Tópico de Atendimento

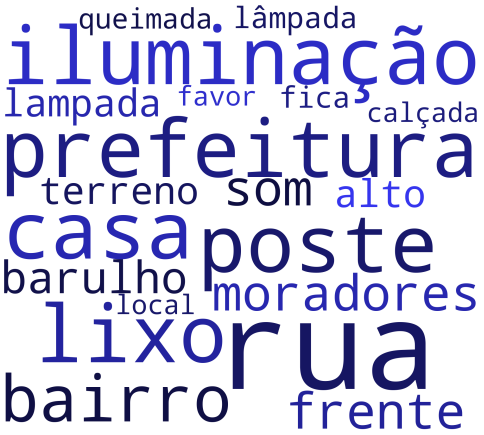

Figura 13. Tópico de Infraestrutura

Figura 14. Tópico de Transporte

Foram gerados sete tópicos dentre os quais selecionou-se quatro muito bem definidos pelo algoritmo. O primeiro tópico refere-se à temática de Telecomunicações, pois nele estão contidos os nomes de prestadora de serviços de telefonia como Vivo, Tim, Net e Oi, além das palavras internet e telefone.

O segundo tópico estabelecido pelo algoritmo LDA trata do assunto referente à reclamação cujo contexto enquadra-se na solicitação do bilhete único, cartão para acessar o transporte público na cidade de São Paulo, e demais assuntos relacionados como passe livre, passagem de estudante e transporte.

O terceiro tópico aproxima-se do tópico reclamações referentes ao tópico Transporte. Nele, estão incluídas as palavras mais frequentes que compõem o tópico, como ônibus, carro, trânsito, motorista. O quarto tópico tendo em vista as palavras que o compõem aproxima-se do tópico Infraestrutura, pois compreende as palavras Iluminação, lixo, lâmpada e queimada que podem ser interpretadas como um indicativo de problemas estruturais.

VII. Conclusão e trabalhos futuros

A partir dos resultados obtidos e das análises então realizadas tornou-se possível afirmar que a criação de modelos a fim de auxiliar no tratamento de reclamações de cidadãos é sim uma tarefa palpável.

Foram exploradas diversas técnicas e abordagens diferentes com a mesma finalidade com o intuito de selecionar as que se saíram relativamente melhores e então, por fim, propôr alternativas eficientes que pudessem vir a trazer melhores resultados para um uso real.

Dentre essas técnicas observou-se que análises realizadas de forma manual, mesmo que com auxílio de soluções como as *Word Clouds*, foram mais acertivas em relação a decisão de qual dos modelos que se sobressaíram em relação aos demais quando comparadas com o simples uso de um índice interno, no caso em questão o Índice *Silhouette*.

É plausível também a observação de que os resultados obtidos tornaram-se mais passíveis de extração de informação na ocasião em que utilizou-se uma técnica mais rebuscada de representação quando comparado com a mudança do algoritmo de agrupamento para algum que fosse mais complexo. Um forte indício dessa afirmação é a melhora expressível dos agrupamentos no momento em que trocou-se a forma de representação das palavras de TFIDF para Word2Vec durante o uso do algoritmo kmeans. Uma melhora a altura não foi alcançada quando trocou-se a técnica de clusterização de kmeans para LDA.

O pressuposto de que clusterizar as palavras dentro de cada um dos agrupamentos de documentos para assim encontrar sub-assuntos aptos a trazerem mais informações valiosas não se concretizou nesse caso. Quando utilizado o *Chinese Restaurant Process* os resultados não agregaram conhecimento o suficiente a ponto de justificar o uso de tal técnica.

Para possíveis trabalhos futuros recomenda-se o teste de novas técnicas de agrupamento para criação de mais modelos a serem comparados com os demais. Além disso, a implementação de um outro índice interno que fosse mais condizente com as análises manuais poderia ser muito útil à medida que automatizaria a comparação de modelos. Por fim, o uso de uma análise fatorial para o agrupamento de categorias no intuito de gerar as super-categorias, usadas para a simplificação da tarefa de clusterização, poderia trazer resultados mais robustos.

Referências

[1] X. Dai, M. Bikdash, and B. C. Meyer, "From social media to public health surveillance: Word embedding based clustering method for twitter classification," *SoutheastCon 2017*, pp. 1–7, 2017.

[2] T. Pratama and A. Purwarianti, "Topic classification and clustering on indonesian complaint tweets for bandung government using supervised and unsupervised learning," in *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, Aug 2017, pp. 1–6.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781

[4] E. Chen, "Introduction to latent dirichlet allocation."

[5] D. Blei, "The chinese restaurant process."

[6] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377042787901257

[7] Repositório de word embeddings do nilc. [Online]. Available: http://nilc.icmc.usp.br/embeddings

[8] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 63–70. [Online]. Available: https://doi.org/10.3115/1118108.1118117