O MODELO PROBABILÍSTICO DE TÓPICOS APLICADO À SEGURANÇA PÚBLICA

Marcio Pereira Basilio
Universidade Federal Fluminense
Rua Passo da Pátria, 156, Campus Praia Vermelha, Bloco D - sala 309
São Domingos, Niterói, RJ, Brasil
CEP: 24.210-240
marciopbasilio@gmail.com

Valdecy Pereira
Universidade Federal Fluminense
Rua Passo da Pátria, 156, Campus Praia Vermelha, Bloco D - sala 309
São Domingos, Niterói, RJ, Brasil
CEP: 24.210-240
valdecypereira@yahoo.com.br

RESUMO

A pesquisa objetivou desenvolver uma metodologia para descoberta de conhecimento em banco de dados dos serviços de atendimento de emergência, com base nos relatos das ocorrências policiais atendidas, com a finalidade de gerar informação para subsidiar os órgãos encarregados de cumprir a lei no planejamento das ações de investigação e combate de ações criminais. A metodologia constituiu-se da utilização de técnicas de mineração de texto, conjugada à técnica LDA (*Latent Dirichlet Allocation*) para obtenção de tópicos sobre a criminalidade. A avaliação e validação dos tópicos foi feita por especialistas da área. Como resultados pode-se relatar que o método utilizado permitiu a identificação dos delitos mais comuns ocorridos no período de 01 de janeiro a 31 de dezembro de 2016, nas áreas estudadas. A análise dos tópicos identificados permitiu reafirmar que os crimes não ocorrem de forma linear em uma determinada localidade, no presente estudo 40% dos delitos identificados na Área Integrada de Segurança Pública nº 5 (AISP 5 Região do Centro da Cidade do Rio de Janeiro), não possuíam correspondência com a Área Integrada de Segurança Pública nº 19 (AISP 19 Bairro de Copacabana), bem como, 33% dos delitos da AISP 19 não foram identificados na AISP 5. Como limitação pode se registrar que os dados coletados representam a dinâmica social dos bairros do centro e da zona sul da cidade do Rio de Janeiro no período específico de janeiro de 2013 a dezembro de 2016. O que implica dizer que os resultados não podem ser generalizados para áreas com características diferentes. A metodologia desenvolvida contribui de forma complementar na identificação de práticas delituosas e suas características a partir dos relatos das ocorrências policiais arquivadas nos bancos de dados dos serviços de emergências. O conhecimento gerado permite aos especialistas dos órgãos encarregados de fazer cumprir a lei avaliar, reformular e construir estratégias diferenciadas para o combate de crimes em determinada localidade. Como implicações sociais pode-se inferir que com a escolha das estratégias adequadas ao combate da criminalidade local, o modelo proposto proporcionará um aumento da sensação de segurança por meio da redução efetiva dos delitos.

Palavra-chave: Modelo Probabilístico de tópicos 1; Mineração de texto; LDA; Crime; Segurança Pública.

ABSTRACT

The research aimed to develop a methodology for the discovery of knowledge in the database of emergency services, based on the reports of police occurrences attended, with the purpose of generating information to support law enforcement agencies in the planning of emergency actions. investigation and combat of criminal actions. The methodology consisted of the use of text mining techniques, combined with the Latent Dirichlet Allocation (LDA) technique to obtain topics on crime. The evaluation and validation of the topics was made by experts in the field. As a result, it can be reported that the method used allowed the identification of the most common crimes occurred from January 1 to December 31, 2016, in the studied areas. The analysis of the identified topics allowed us to reaffirm that the crimes do not occur linearly in each locality. In the present study 40% of the crimes identified in the Integrated Area of Public Security No. 5 (AISP 5 Region of the City Center of Rio de Janeiro), corresponded with the Integrated Public Safety Area No. 19 (AISP 19 Bairro de Copacabana), as well as 33% of the crimes of AISP 19 were not identified in AISP 5. As a limitation it can be registered that the collected data represent the social dynamics. from the downtown and southern districts of the city of Rio de Janeiro in the specific period from January 2013 to December 2016. This implies that the results cannot be generalized to areas with different characteristics. The developed methodology contributes in a complementary way in the identification of criminal practices and their characteristics from the reports of the police occurrences filed in the emergency services databases. The knowledge generated enables law enforcement specialists to evaluate, reformulate and build different strategies for combating crime in each locality. As social implications it can be inferred that with the choice of appropriate strategies to combat local crime, the proposed model will provide an increased sense of security through the effective reduction of crimes.

Keywords: Probabilistic topic model 1; Text mining; LDA; Crime; Public security.

Como Citar:
BASILIO, Marcio; PEREIRA, Valdecy. O modelo probabilístico de tópicos aplicado à Segurança Pública. *In*: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 19., 2019, Rio de Janeiro, RJ. **Anais** […]. Rio de Janeiro: Centro de Análises de Sistemas Navais, 2019.

1. INTRODUÇÃO

Ao longo dos últimos anos inúmeras pesquisas foram desenvolvidas em torno do tema, buscando compreender as causas relacionadas à incidência criminal e suas variações (AGNEW, 2016; SHERMAN, GARTIN e BUERGER, 1989; WEISBURD e ECK, 2004; HABERMAN, 2017), bem como, identificar práticas e estratégias de combate ao crime. Há uma discussão sob a efetividade das estratégias preventivas e repressivas adotadas por Estados, no controle da criminalidade (SHERMAN, MACKENZIE, *et al.*, 1998; BRAGA, 2001). Sherman, et al. (1998) em seu relatório de pesquisa avaliaram as estratégias que foram utilizadas no contexto norte-americano sob a ótica de sua eficácia em função dos resultados obtidos. Outros estudos trataram da analise criminal e argumentam que os crimes não ocorrem de forma uniforme nas cidades e que existem agrupamentos significativos de delitos em lugares que são denominadas de hot spot. Vários pesquisadores argumentam que os crimes podem ser reduzidos de forma eficiente se as estratégias fossem direcionadas aos lugares de maior concentração criminal (BRAGA, 2005; SHERMAN e WEISBURD, 1995).

De uma forma geral os estudos identificam quatro tipos de estratégias utilizadas pelas agencias encarregadas de aplicar a lei, em diversos contextos, que são:

- *Standard Model of Policing* (BAYLEY, 1994);
- *Community Policing* (SKOLNICK e BAYLEY, 1986);
- *Problem-Oriented Policing* (GOLDSTEIN, 1990); e
- *Hot Spots Policing* (BRAGA, 2001).

Todavia, o enorme volume de dados, oriundos do registro do relato das circunstancias; local; características físicas; dinâmica do fato delituoso, que são armazenados diariamente pelos serviços de emergências no mundo inteiro são uma fonte de dados não estruturada que podem fornecer informações que subsidiem o planejamento das atividades policiais, que contribuem para indicar a estratégia adequada em determinada localidade, e nas investigações criminais. Desta forma, esta pesquisa procurou resposta para seguinte questão: Como os relatos de atendimento das ocorrências policiais realizado pelos serviços de emergências, podem contribuir para escolha da estratégia de combate ao crime em uma determinada localidade?

Neste sentido, a pesquisa teve como objetivo principal desenvolver um modelo de ordenação das estratégias de policiamento em função dos delitos recorrentes em uma determinada localidade. A metodologia desenvolvida para resolução do problema integrou técnicas de mineração de texto, por meio da utilização da técnica *Latent Dirichlet Allocation* (LDA) (BLEI, NG e JORDAN, 2003) para obtenção de tópicos sobre a criminalidade, que com os resultados do LDA criará um ranking das estratégias de combate ao crime nas localidades estudadas. A aplicação deu-se na região metropolitana da capital do Estado do Rio de Janeiro, Brasil, em colaboração com a agência local encarregada pela aplicação da lei. Como resultado do modelo desenvolvido, foram identificados dez tópicos, que após o processo de validação por especialistas, foram rotulados como os delitos com maior insurgência nas áreas estudada.

2. REVISÃO TEÓRICA

2.1. UMA VISÃO SOBRE MINERAÇÃO DE TEXTOS

A mineração de texto é o processo de descobrir informações importantes e recursos de dados textuais (CHEN, LIU e HO, 2013). Como relatado em (MORAIS e AMBRÓSIO, 2007) a mineração de textos tem sua origem relacionada a área de *Knowledge Discovery from Text* - KDT, tendo seus processos sido descritos pela primeira vez em (FELDMAN e DAGAN, 1995), descrevendo uma forma de extrair informações a partir de coleções de texto dos mais variados tipos. Atualmente, mineração de textos pode ser considerada sinônimo de descoberta de conhecimento em textos. As principais contribuições desta área estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e à melhor compreensão de textos disponíveis em documentos. Textos estes que podem estar representados das mais diversas formas, dentre elas: e-mails; arquivos em diferentes formatos (pdf, doc, txt, por exemplo); páginas Web; campos textuais em bancos de dados; textos eletrônicos digitalizados a partir de papéis. A mineração de textos estruturados é encontrada em campos do conhecimento tais como: bibliometria, cientometria, informetria, midiametria, museometria e webmetria (CAPUANO, 2009).

Recentemente, a mineração de texto tornou-se uma importante área de pesquisa. A mineração de texto é um campo interdisciplinar de várias tecnologias, incluindo bancos de dados, mineração de dados, recuperação de informações, linguística, estatística dentre outros. Como a maior parte do conhecimento e da história humana são armazenadas em documentos que contêm texto, os textos são um rico depósito de informações preciosas. Dependendo do tipo de documento, diferentes partes de informações valiosas são ocultadas (CHEN, LIU e HO, 2013). A importância da utilização da técnica de mineração de texto pode ser constatada por meio das diversas aplicações e métodos que foram desenvolvidos, conforme afirmam (ALWIDIAN , BANI-SALAMEH e ALSLAITY, 2015). Por exemplo: news categorization; patent retrieval; e-mail security; scientific document retrieval; theme detection; document sentiment analysis; authorship identification; document summarization; e search engines

2.2. MODELOS PROBABILÍSTICOS DE TÓPICOS

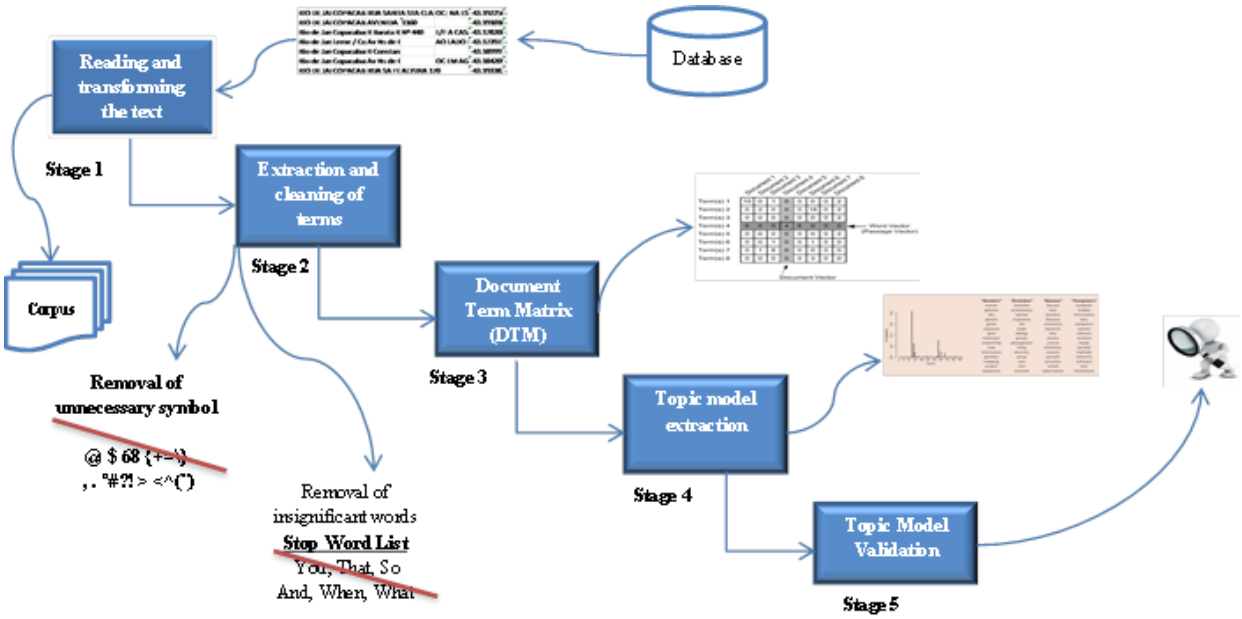
A exploração de grandes volumes de dados é simplificada pelos modelos probabilísticos na descoberta dos tópicos. Os tópicos são estruturas com valor semântico e que, no contexto de mineração de texto, formam grupos de palavras que frequentemente ocorrem juntas. Esses grupos de palavras quando analisados, dão indícios a um tema ou assunto que ocorre em um subconjunto de documentos. A expressão tópico, conforme (FALEIROS e LOPES, 2016), é usada levando-se em conta que o assunto tratado em uma coleção de documentos é extraído automaticamente, ou seja, tópico é definido como um conjunto de palavras que frequentemente ocorrem em documentos semanticamente relacionados.

O *Latent Dirichlet Allocation* (LDA) é um modelo probabilístico generativo para coleções de dados discretos como um conjunto de documentos (corpus). Um modelo generativo é aquele que aleatoriamente gera os dados a partir das variáveis latentes. Nesse modelo, as variáveis observáveis são os termos de cada documento e as variáveis não observáveis são as distribuições de cada tópico (BLEI, NG e JORDAN, 2003). Os parâmetros das distribuições de tópicos, conhecidos como hiper-parâmetros, são dados *a priori* no modelo. A distribuição utilizada para amostrar a distribuição de tópicos é a distribuição de *Dirichlet*. No processo generativo, o resultado da amostragem da *Dirichlet* é usado para alocar as palavras de diferentes tópicos e que preencherão os documentos. Assim, pode-se perceber o significado do nome *Latent Dirichlet Allocation*, que expressa a intenção do modelo de alocar os tópicos latentes que são distribuídos obedecendo a distribuição de *Dirichlet*. O LDA é baseado na intuição que cada documento contém palavras de múltiplos tópicos; a proporção de cada tópico em cada documento é diferente, mas os mesmos tópicos são os mesmos para todos os documentos.

3. METODOLOGIA

A metodologia desenvolvida neste trabalho, ilustrada na Figura 1, se subdivide em cinco estágios. O primeiro foi denominado de “Leitura e transformação do texto”. Neste estágio serão identificados e extraídos os campos do banco de dados analisado. Em seguida serão carregados para o computador e individualizados na forma de arquivos textos. Estes arquivos constituirão o *corpus*. O segundo estágio denominado de “Extração e limpeza dos termos”, será responsável pela decomposição do corpus em termos (tokenização), neste processo serão eliminados os símbolos e caracteres de controle de arquivo ou de formatação, bem como os sinais de pontuação, números, datas. Os múltiplos espaços serão reduzidos a espaços simples. Em seguida inicia-se o processo de limpeza para a retirada das *stops words*, que são compostas por: preposições, artigos, advérbios, números, pronomes, conjunções, interjeição e pontuação. O terceiro estágio será responsável por categorizar os termos e associando-os às respectivas frequências de ocorrência no corpus analisado, possibilitando inferência sobre suas proximidades, distâncias e termos relacionados. No quarto estágio, em função da constituição do *Document Term Matrix* (DTM), será realizado o processo de identificação dos *topic model*, com a utilização do *Latent Dirichlet Allocation* (LDA) com *Gibbs Sampling* Colapsado. Por fim, no quinto estágio denominado de “Validação dos *Topic Model*” serão criados questionários com os resultados da lista de termos constituinte de cada tópico para serem submetidos a um grupo de especialistas da área, para a definição do rótulo de cada tópico.

Figura 1Ilustração dos estágios do processo de obtenção do modelo de tópicos



Fonte: (BASILIO, et al., 2019)

3.1. UNIVERSO E AMOSTRA

O universo da pesquisa foi os registros de atendimentos de chamadas de emergência, realizados pelo serviço de 190 da Polícia Militar do Estado do Rio de Janeiro, no período compreendido entre 01 de janeiro de 2013 a 31 de dezembro de 2016. Ocorridos na região metropolitana da Cidade do Rio de Janeiro. Neste período foram registrados pelo sistema 29.627.559 de chamadas.

3.2. COLETA E TRATAMENTO DOS DADOS

A coleta de dados referente ao serviço de atendimento 190, foi realizada, em colaboração com a agência local encarregada de fazer cumprir a lei (Polícia Militar do Estado do Rio de Janeiro), a qual disponibilizou os arquivos relativos aos registros mensais dos atendimentos efetuados no período de 01 de janeiro de 2013 a 31 de dezembro de 2016. O tratamento dos dados será feito com a utilização do *Software* R.

4. APLICAÇÃO, ANÁLISE E DISCUSSÃO DOS RESULTADOS

Para fins de aplicação da metodologia, a pesquisa foi desenvolvida na Cidade do Rio de Janeiro, no Estado do RJ no Brasil. O estudo foi desenvolvido com base no serviço de atendimento de emergência policial denominado “190”. Sendo assim, foi selecionado o campo “Batalhão” referente à base de dados “190”, e de forma aleatória foram escolhidas duas áreas integradas de segurança pública(AISP), sendo resultante deste processo as AISP 5 (5º BPM)e 19 (19º BPM), corresponde as áreas do Centro da cidade e Copacabana respectivamente.

Após a seleção das áreas de policiamento que seriam analisadas, iniciou-se a aplicação da metodologia ilustrada na Fig. 1. No primeiro estágio foi feita a identificação dos registros correspondentes a categoria "ocorrência" relacionadas ao 5º BPM e ao 19º BPM, referentes ao ano de 2016. A extração resultou, no caso do 5º BPM em um total de 24.286 arquivos de texto. Em relação ao 19º BPM foram extraídos 14.374 arquivos.

Em seguida, no segundo estágio foram constituído dois *corpus*: o primeiro, referente ao 5º BPM, composto por 24.286 elementos. O segundo, em relação ao 19º BPM com 14374 elementos. Finalizada a criação dos *corpus*, utilizou-se o *software* R para eliminação dos espaço em brancos, caracteres especiais, pontuação, números, acentos e transformação das vogais e consoantes maiúsculos em minúsculos.

Terminado o processo de tokenização, foi realizado o processo de limpeza dos textos constituintes dos *corpus*, removendo-se as *stop words*. Na presente pesquisa, foram removidas os seguintes morfemas lexicais: preposições; artigos; advérbios; números; pronomes; conjunções; interjeição e pontuações. A remoção foi realizada com os morfemas lexicais em inglês (183 itens) e em português (742 itens). Além destas classes de palavras, foram retirados morfemas como siglas (1633 itens); fonemas sem sentidos; alfabeto militar; designação de meses; nomenclatura dos postos e graduações utilizados pelas instituições policiais; abreviaturas de unidades policiais.

Após a remoção dos morfemas lexicais, foram feitas as identificações e substituições de siglas e morfemas por sinónimos, que simplificam a análise do material. Após a limpeza do corpus, o próximo processo seria a Stemming. Todavia, o presente corpus, possui características em que a redução dos morfemas a seus radicais traria prejuízo a análise e constituição dos tópicos, pois as desinências (morfemas flexionais), afixo (morfemas derivacionais) e a vogal temática, como assevera (CUNHA, 2001), são importantes para diferenciar o agente ativo do agente ativo da ação; ou se a ação foi finalizada ou se ainda estava em andamento; se os agentes das ações são do gênero masculino ou feminino. Cabe ressaltar, que o corpus desta pesquisa refere-se a ações delituosas atendidas pelas instituições policiais, e a redução dos morfemas a seus radicais traria perda de significado para a análise. Neste sentido, decidiu-se não aplicar esta etapa a presente pesquisa.

No terceiro estágio, após a execução das etapas de pré-processamento foi gerada a DTM. Inicialmente o *corpus* do 19º BPM que foi reduzido de 14374 elementos para 10427. Isto ocorre em função da não alocação de todos os termos em todos os documentos analisados. O resultado é a geração de uma matriz com inúmeros espaços vazios, o que os especialistas denominam de matrizes esparsas (DAVIS e HU, 2011; DUFF, GRIMES e LEWIS, 1989). Após o procedimento de redução de espaços vazios na matriz, a uma taxa de 0.9999, chegou-se a um DTM com 4918 elementos. Em relação ao corpus do 5º BPM, iniciou-se com 24286 elementos após o pré-processamento chegou-se à 14786 elementos. Realizando-se o procedimento de redução de espaços vazios na matriz, a uma taxa de 0.999, obteve-se um DTM com 4918 elementos.

No quarto estágio, iniciou-se o processamento dos dados aplicando o método *Latent Dirichlet Allocation* (LDA) com Gibbs sampling, tendo sido utilizado o *software* R, para obtenção dos tópicos. Desta forma, foram obtidos os tópicos referentes ao 5º e 19º BPM, conforme ilustrado nas Fig. 2 e 3. Como principais saídas do modelo LDA tem-se os vetores-tópico, que são as distribuições sobre os termos do vocabulário fixo que caracteriza cada tópico, e os vetores-documentos, que são as distribuições de frequência relativa da

ocorrência de cada tópico para um dado documento, conforme apresentado nas Tabelas 9 e 10. A partir destas informações foram produzidas informações sobre a participação de cada tópico na amostra estudada. Na amostra referente ao 5º BPM, pode-se observar que os tópicos 5, 8, 1, e 3 possuem as maiores frequências relativas de ocorrência entre os documentos analisados, conforme representado na Tabela 9 e ilustrado na Fig. 14. Em relação a amostra do 19º BPM, destacam-se os tópicos 9, 1, 2, 3, e 4, conforme representado na Tabela 10 e ilustrado na Fig. 15. Cabe ressaltar, que a disposição dos tópicos nas figuras não possuem nenhum tipo de ranking entre os mesmos.

Figura 2 Lista dos tópicos do 5º BPM em 2016

Tópico 1	• entorpecentes armados moradores usando vendendo uso ocorre
Tópico 2	• tentando prédio frente porta loja maiores ligacao
Tópico 3	• arma ameacando pessoas fogo faca homem maiores
Tópico 4	• loja estabelecimento desentendimento atrito porta funcionarios tentando
Tópico 5	• transeuntes menores roubando roubos individuos cerca roubo
Tópico 6	• moto vitima veiculo roubado carro colisao hospital
Tópico 7	• residencia mulher agredindo agredida agrediu marido vizinha
Tópico 8	• armado individuo celular magro mochila armados pertences
Tópico 9	• disparo alarme agencia banco imagens equipe brasil
Tópico 10	• som alto proveniente bar festa vizinho incomodando

Fonte: Elaborado pelos autores

Figura 3 Lista dos tópicos do 19º BPM em 2016

Tópico 1	• armados comunidade disparo_arma_fogo radio telefone ocorre morador
Tópico 2	• som alto proveniente bar incomodando barulho festa
Tópico 3	• disparo alarme banco agencia porta imagens equipe
Tópico 4	• residencia vizinho agredida agrediu gritando marido mae
Tópico 5	• loja atrito estabelecimento desentendimento cliente verbal pertences
Tópico 6	• entorpecentes moradores frente usando uso consumindo pessoas
Tópico 7	• ameaçando pessoas tentando faca armado agressao morador
Tópico 8	• onibus sentido bicicleta suspeita coletivo direcao atitude
Tópico 9	• transeuntes cerca menores grupo individuos roubos alertados
Tópico 10	• residencia mulher agredindo homem agredida agrediu residencial

Fonte: Elaborado pelos autores

No quinto estágio, após a obtenção dos tópicos de cada área de policiamento por meio do método LDA, buscou-se validar as informações obtidas junto à especialistas que atuam diretamente no atendimento e controle das ocorrências policiais. Neste sentido, construiu-se um questionário contendo vinte e cinco questões. As cinco primeiras visavam construir um perfil dos especialistas. Da sexta a vigésima quinta questões foram dispostas, em colunas, as sete palavras de cada tópico, acrescida de mais uma palavra diferente do contexto. A inclusão, desta palavra, objetivava testar a coerência do conjunto das sete palavras, buscando conhecer se de fato representavam um tópico real.

4.1. IDENTIFICAÇÃO DAS PALAVRAS DE CONTROLE

Após caracterização da amostra, passou a analisar a coerência do conjunto de palavras de cada tópico. Nesta etapa, a partir do cálculo de frequência das respostas assinaladas por cada respondente, observa-se que em 95% dos casos, as palavras inseridas no conjunto estudado foram identificadas, indicando que o as palavras que compunham os tópicos apresentavam coerência e os representava. Em média o percentual de identificação de cada palavra de controle foi de 90,2%, com desvio padrão de 11,95% , sendo o mínimo de 42 e o máximo de 97%. O caso em que houve uma discordância, ocorreu no tópico 6 referente a área do 5º BPM, analisando este caso, observou-se que 46% indicaram a palavra “BRASIL” e 42% assinalaram a palavra de controle. No caso especifico a palavra “BRASIL” corresponde ao nome de uma instituição bancária, e o tópico em questão apontava para ocorrência do sistema bancário.

4.2. IDENTIFICAÇÃO DOS RÓTULOS DOS TÓPICOS

Nesta fase, foi utilizada estatística descritiva para identificar os rótulos atribuídos pelos respondentes da pesquisa ao conjunto de palavras de cada tópico.

4.2.1. Rótulos dos tópicos relativos à área do 19º BPM

Tabela 1 Validação dos rótulos correspondentes aos tópicos 1-10 do 19º BPM

Tópico	Denominação	Frequência	Porcentagem	Porcentagem valida
1	Roubo de rua	86	86	86
2	Roubo a estabelecimento financeiro	60	60	60
3	Violência doméstica	61	61	61
4	Roubo	51	51	51
5	Perturbação do sossego	76	76	76
6	Ameaça	41	41	41
7	Roubo de veículo	73	73	73
8	Violência doméstica	52	52	52
9	Disparo de arma de fogo	60	60	60
10	Uso de entorpecentes	87	87	87

Fonte: Elaborado pelos autores

Os tópicos numerados de 1 a 10 do questionário correspondem ao levantamento realizado na área de atuação do 19º BPM. A Tabela 1 apresenta as frequências identificadas dos rótulos atribuídos a cada tópico pelos respondentes. Cabe ressaltar que, em consequência da variedade de rótulos, foi realizado categorização dos rótulos similares ou que foram estratificados, interessando para a pesquisa a tipificação dos delitos de forma geral.

4.2.2. Rótulos dos tópicos relativos a área do 5º BPM

Tabela 2 Validação dos rótulos correspondentes aos tópicos 1-10 do 5º BPM

Tópico	Denominação	Frequência	Porcentagem	Porcentagem valida
1	Perturbação do sossego	94	94	94
2	Tráfico de drogas	68	68	68
3	Ameaça	45	45	45
4	Roubo	42	42	42
5	Acidente de trânsito	78	78	78
6	Disparo de alarme bancário	43	43	43
	Roubo a estabelecimento financeiro	48	48	48
7	Violência Doméstica	73	73	73
8	Roubo a veículo	36	36	36
9	Roubo de rua	89	89	89
10	Individuo armado	49	49	49

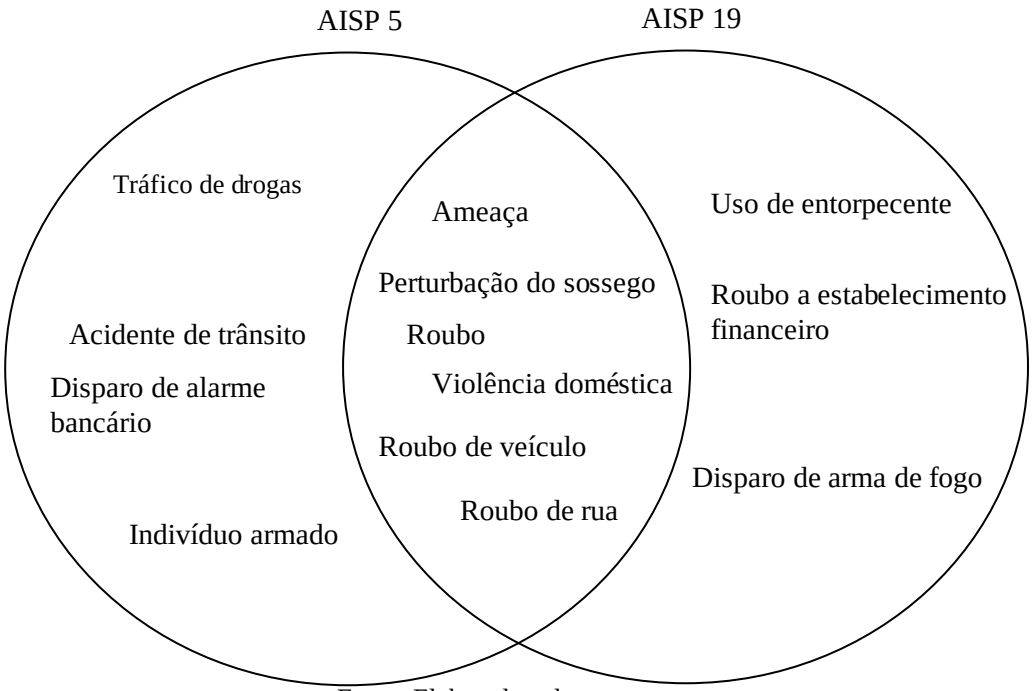
Fonte: Elaborado pelos autores

Os tópicos numerados de 1 a 10 do questionário correspondem ao levantamento realizado na área de atuação do 5º BPM. A Tabela 2 apresenta as frequências identificadas dos rótulos atribuídos a cada tópico pelos respondentes. Cabe ressaltar que, em consequência da variedade de rótulos, foi realizado categorização dos rótulos similares ou que foram estratificados, interessando para a pesquisa a tipificação dos delitos de forma geral.

4.3. IDENTIFICAÇÃO FINAL DOS TOPIC DE CADA ÁREA DE POLICIAMENTO

Após a validação dos tópicos pelos especialistas, decidiu-se por nomear os tópicos pelos rótulos com maior frequência na amostra. Desta forma, a figura 4 ilustra os tópicos validados na Tabela 1 e 2, bem como, ilustra a interseção das demandas comuns entre as duas AIPSs.

Figura 4 Representação gráfica da interseção das demandas das áreas integradas de segurança pública pesquisadas



Fonte: Elaborado pelos autores

Após a obtenção dos tópicos por meio do LDA, foi utilizado o *package* "leaflet" em R, para ilustrar os tópicos distribuídos nas áreas estudadas. O processo de identificação dos tópicos resulta na associação dos tópicos aos documentos analisados. A partir deste ponto, foi possível recuperar os dados de localização gerados no momento do atendimento serviço de emergência 190. Com dados de latitude e longitude das ocorrências atendidas e analisadas neste estudo e sua associação aos tópicos foi possível gerar as Figuras 5 a 8.

As Figuras 5 a 8 ilustram a localização de 24.020 ocorrências policiais relacionadas aos dez temas identificados na área de policiamento do 5º BPM. A quantidade de ocorrências corresponde ao serviço realizado no período de 1º de janeiro de 2016 a 31 de dezembro de 2016.

As Figuras 5 e 6 correspondem a uma visão geral da localização espacial da área onde o estudo foi desenvolvido. Nessas figuras, o círculo com a numeração corresponde aos clusters dos dez tópicos. Na Figura 7, o leitor é apresentado á uma visão mais detalhada, em que é possível observar a existência de vários clusters referentes aos dez tópicos estudados. A numeração dentro dos círculos coloridos corresponde ao número de ocorrências atendidas em uma área delimitada. A Figura 8 mostra uma vista explodida de um determinado cluster, onde é possível identificar a composição do cluster pelas demandas atendidas.

Figura 5 Visão geral da área de estudo
Fonte: Criado com o package "Leaflet" em R.

Figura 6 Visão aproximada da área de estudo correspondente ao Rio de Janeiro, Brasil.
Fonte: Criado com o package "Leaflet" em R.

Figura 7 Visão aproximada da área de policiamento do 5º BPM, localizada no Centro da Cidade do Rio de Janeiro, Brasil.
Fonte: Criado com o package "Leaflet" em R.
Nota: Nesta figura podemos observar um conjunto de clusters correspondentes aos dez tópicos observados na área do 5º BPM. Cada cluster contém um número referente ao número de chamadas feitas durante o período considerado.

Figura 8 Visão explodida de um cluster correspondente a diversos tópicos, dentro da área de policiamento do 5º BPM, localizado no Centro da Cidade do Rio de Janeiro, Brasil.
Fonte: Criado com o package "Leaflet" em R.
Nota: Os tópicos são identificados pelas seguintes cores: Tópico 1 (verde); Tópico 2 (laranja); Tópico 3 (preto); Tópico 4 (azul); Tópico 5 (rosa); Tópico 6 (bege); Tópico 7 (cinza); Tópico 8 (azul escuro); Tópico 9 (verde escuro); e Tópico 10 (vermelho).

Restaurando a ideia central do LDA, o qual assume que cada documento num corpus é gerado por uma mistura de diferentes proporções de um número limitado de tópicos, e cada tópico contribui com várias palavras associadas a ele, pode-se inferir que as figuras 2 e 3 revelam a estrutura latente dos tópicos do corpus analisado, que é composto por documentos originados a partir do campo inerente ao registro dos relatos das ocorrências, feita pelo serviço de atendimento 190. Sendo K, o número fixo de tópicos, na pesquisa K=10, como registrado no Apêndice A, foram extraídos dez tópicos para cada área de policiamento estudada. Os termos associados a cada tópico permitem aos especialistas em segurança pública inferirem, sobre os elementos pertinentes aos delitos que foram cometidos em cada uma das áreas pesquisadas, como por exemplo, o tópico 10 da área do 5º BPM, os termos associados sugerem que há um problema recorrente originado pelo desrespeito as normas de posturas municipais, no que tangem ao nível de ruído autorizado. Associado aos locais de emissão de som alto pode-se inferir também, que são locais propensos a ocorrência de outros delitos associados como: lesões corporais, rixas, e até homicídios. Desta forma, as informações recuperadas por meio da extração dos tópicos podem de forma complementar auxiliar o planejamento operacional, aplicação de recursos materiais e humanos na prevenção dos delitos. Por outro lado, o processo de validação dos tópicos reforçou a ideia latente que os termos associados a cada tópico referiam-se a um delito específico. Neste sentido, o processo de validação feitos por meio dos especialistas em segurança pública permitiu a rotulação dos tópicos, sendo validado dez tópicos para a área analisada referente ao 5º BPM, e nove tópicos inerente a área do 19º BPM, conforme ilustrado na Figuras 4. A rotulação dos tópicos permite a identificação das demandas recorrentes em relação as ocorrências policiais. Analisando a figura 4 constata-se que 40% dos delitos identificados por meio dos tópicos na área do 5º BPM não foram observadas na área do 19º BPM. Da mesma forma, revela que 33% da demanda do 19º BPM não foram observadas na área de atuação do 5º BPM. Estas observações nos levam a inferir que as demandas são diferentes em cada área de policiamento, reforçando o argumento de que o crime não se comporta de forma linear. Esta diferença conduz ao raciocínio que para cada área de policiamento deva ser aplicada um conjunto de estratégias especificas, bem como, uma aplicação de recursos diferenciada. Não obstante a isto, cabe relatar que existe uma área comum de interseção entre as localidades, que na Figura 4 pode-se dizer que são: ameaça; perturbação ao sossego; roubo; roubo de veículo; roubo de rua; e violência doméstica.

5. CONSIDERAÇÃO FINAL

O método desenvolvido na pesquisa, consistiu inicialmente na extração de relatos de atendimento das ocorrências policiais do banco de dados do serviço de emergência, que após o processo de mineração de texto, permitiu a utilização do *Latent Dirichlet Allocation* com Gibbs Sampling Colapsado culminando com a extração dos *topic model* das áreas de policiamento estudadas. Foram identificados dez *topic model* de cada área pesquisa. Este resultado auxilia os especialistas na identificação os termos associados a cada tópico. Tal procedimento autoriza a inferência sobre características de cada delito, o que contribui subsidiariamente na compreensão da dinâmica de cada delito, permitindo os ajustes necessários no planejamento do combate ao crime, na escolha da estratégia mais adequada em uma determinada área de policiamento. Outra contribuição foi o processo de validação dos *tópic model* com a utilização de especialistas. A utilização dos especialistas foi fundamental para associação dos termos de cada tópico com uma tipificação criminal.

Com a rotulação dos tópicos, identificaram-se os tipos latentes de demandas do serviço de emergência em cada uma das áreas estudadas. Neste sentido, constatou-se que 40% das

demandas da AISP 5 não foram identificadas na AISP 19. Da mesma forma, 33% da demanda ocorrida na AISP 19 não ocorriam na AISP 5. Sendo assim, este resultado corrobora com o argumento de que o crime não ocorre de forma linear, necessitando, com isso, de estratégias diferenciadas para o seu combate.

REFERÊNCIA BIBLIOGRÁFICA

[1] AGNEW,. A theory of crime resistance and susceptibility. **Criminology**, 54, n. 2, 2016. 181-211. doi: 10.1111/1745-9125.12104.

[2] ALWIDIAN , S. A.; BANI-SALAMEH, H. A.; ALSLAITY, A. N. Text data mining: A proposed framework and future perspectives. **International Journal of Business Information Systems**, 18, n. 2, 2015. 127-140.

[3] BAYLEY, D. H. **Police for the future**. New York: Oxford University Press, 1994.

[4] BASILIO, M. B.; PEREIRA, V.; BRUM, G. Identification of operational demand in law enforcement agencies. **Data Technologies and Applications**, V. 53 n. 3, pp. 333-372, 2019. https://doi.org/10.1108/DTA-12-2018-0109

[5] BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, 3, 2003. 993-1022.

[6] BRAGA, A. The Effects of Hot Spots Policing on Crime. **The Annals of the American Academy of Political and Social Sciense**, 578, n. 1, 2001. 104-125.

[7] BRAGA, A. A. Hot spots policing and crime prevention: A systematic review of randomized controlled trials. **Journal of Experimental Criminology**, n. 1, 2005. 317–342.

[8] CAPUANO, A. The cognitive power of artificial neural networks model ART1 for information retrieval. **Ciência da Informação**, 38, n. 1, 2009. 9-30. https://dx.doi.org/10.1590/S0100-19652009000100001.

[9] CHEN, Y.-L.; LIU, Y.-H.; HO, W.-L. A text mining approach to assist the general public in the retrieval of legal documents. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY**, 64, n. 2, February 2013. 280–290.DOI:10.1002/asi.22767.

[10] CUNHA, C. **Nova gramática do português contemporâneo**. 3. ed. Rio de Janeiro: Nova Fronteira, 2001.

[11] DAVIS, T. A.; HU, Y. The University of Florida Sparse Matrix Collection. **ACM Transactions on Mathematical Software**, 38, n. 1, 2011.

[12] DUFF, I. S.; GRIMES, R. G.; LEWIS, J. G. Sparse matrix test problems. **ACM Transactions on Mathematical Software (TOMS)**, 15, n. 1, 1989. 1-14. Doi:10.1145/62038.62043.

[13] FALEIROS, T. D. P.; LOPES, A. D. A. **MODELOS PROBABILÍSTICOS DE TÓPICOS: DESVENDANDO O LATENT DIRICHLET ALLOCATION**. Universidade de São Paulo. São Carlos, p. 59. 2016. (ISSN 0103-2569).

[14] FELDMAN, ; DAGAN, I. **Knowledge Discovery in Textual Databases (KDT)**. THE FIRST INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. Montreal: [s.n.]. 1995. p. 112-117.

[15] GOLDSTEIN, H. **Problem-oriented policing**. New York: McGraw-Hill, 1990.

[16] HABERMAN, C. P. Overlapping Hot Spots? Examination of the Spatial Heterogeneity of Hot Spots of Different Crime Types. **Criminology and Public Policy**, 16, n. 2, May 2017. 633-660. https://doi.org/10.1111/1745-9133.12303.

[17] MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de Textos**. Universidade Federal de Goiás. [S.l.], p. 29. 2007. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf.

[18] SHERMAN, L. W.; GARTIN, P. R.; BUERGER, M. E. Hot Spots of predatory crime: routine activities and the criminology of place. **Criminology**, 27, n. 1, 1989. 27–56. DOI:10.1111/j.1745-9125.1989.tb00862.x.

[19] SHERMAN, L.; WEISBURD, D. General deterrent effects of police patrol in crime hot spots: A randomized controlled trial. **Justice Quarterly**, 12, 1995. 625-648.

[20] SHERMAN, W. et al. **Preventing crime: what works, what doesn´t, what´s promising**. [S.l.]. 1998.

[21] SKOLNICK, J. H.; BAYLEY, D. H. **The new blue line:** Police innovation in six American cities. New York: Free Press, 1986.

[22] WEISBURD, ; ECK, E. What Can Police Do to Reduce Crime, Disorder, and Fear? **The Annals of the American Academy of Political and Social Sciense**, 593, n. 1, 2004. 42-65.