Técnicas de Processamento de Linguagem Natural em Denúncias Criminais: Automatização e Classificação de Texto em Português Coloquial

Techniques of Natural Language Processing in Criminal Reports: Automation and Classification of Text in Colloquial Portuguese

Camila Gusmão[1], Karla Figueiredo[1], Walkir A.T. Brito[2]

[1] Instituto de Matemática e Estatística – Departamento de Ciência da Computação
Universidade do Estado do Rio de Janeiro (UERJ) – Rio de Janeiro – RJ – Brasil

[2] Disque-Denúncia – Rio de Janeiro – RJ – Brasil

camilaeleuteriogusmao@gmail.com, karlafigueiredo@ime.uerj.br,
walkir.brito@disquedenuncia.org.br

Resumo. *Este artigo apresenta a investigação de Técnicas de Processamento de Linguagem Natural (PLN) em Denúncias Criminais, provenientes do aplicativo do serviço do Disque Denúncia RJ para smartphone. Nele é apresentado o processo de automatização, avaliando e classificando as denúncias, objetivando reduzir o tempo de análise do conteúdo das mensagens, que possui, como principal desafio, textos escritos em linguagem muito informal, contendo muitos erros morfossintáticos. Para alcançar tais objetivos foi necessária uma investigação de técnicas de pré-processamento visando melhorar a acurácia da classificação, que foi realizada por Support Vector Machine (SVM). Os resultados encontrados são bastante promissores para o tipo de textos de denúncias, atingindo uma precisão de 76,11%.*

Palavras-chave: *Mineração de Texto, Máquina de Aprendizado, Português Coloquial, Denúncias Criminais*

Abstract. *This article presents the investigation of Natural Language Processing Techniques (PLN) in Criminal reports, from the application of the Disque Denúncia RJ service for smartphone. It presents the automation process, evaluating and classifying reports, aiming to reduce the time of analysis of the content of messages, which has, as its main challenge, texts written in very informal language, containing many morphosyntactic errors. To achieve these goals, an investigation of preprocessing techniques was necessary to improve the accuracy of the classification, which was performed by a Support Vector Machine (SVM). The results found are very promising for the type of denunciation texts, reaching an accuracy of 76.11%.*

Key-words: *Text Mining, Machine Learning, Portuguese Colloquial, Criminal Reports*

1. Introdução

A insegurança tem sido uma grande preocupação para os brasileiros, e no Rio de Janeiro a sensação de insegurança se torna ainda mais alarmante. Em 2019, a região metropolitana do estado do Rio de Janeiro registrou 7.365 tiroteios e disparos de arma de fogo, uma média de 20 tiroteios por dia, que deixaram 2.876 baleados [Carta Capital, 2020]. O Rio de Janeiro conta com o Disque Denúncia (DD), uma central de atendimentos criada para ajudar as polícias Federal, Civil e Militar no esclarecimento de crimes e delitos. Feitas de formas anônimas pela população. Uma característica importante nesse cenário é que devido ao medo de represálias os denunciantes de regiões dominadas pelo tráfico ou por áreas de milícias, normalmente não utilizam os canais formais de denúncias, como os da polícia militar e civil. Desta forma, as denúncias tornam-se uma fonte especialmente importante nessa dinâmica, pois inúmeras categorias de crimes não são denunciadas diretamente ao Estado, seja por medo, ameaça ou até a própria descrença nas autoridades. É nesta circunstância que o DD tem um papel fundamental, possibilitando denúncias de forma anônima em seus canais de recebimento e deste modo continuar contribuindo com a segurança pública. Assim, as denúncias ao DD ocorrem porque seu anonimato é sempre garantido e precisa ser preservado para manter sua efetividade. Assim, a denúncia torna-se, a arma do cidadão, em conduzir o processo investigativo a seu favor [Mendonça, 2007].

Para que tais denúncias sejam mais rapidamente processadas, o desenvolvimento de ferramentas baseadas em *Machine Learning*, que automatize o processo de análise e classificação destas, pode ajudar a viabilizar as investigações criminais para a solução de crimes em tempo hábil. Assim, tal ferramenta torna-se importante e desejável para reduzir o tempo de análises feitas de forma manual, sem deixar de priorizar o desempenho da classificação, para que estas informações sejam adequadamente encaminhadas e avaliadas por entes da Segurança Pública.

Assim, para atingir esses objetivos, este trabalho investigou e adaptou o uso de técnicas de PLN aplicadas aos dados extraídos dos aplicativos para smartphone do DD, em português coloquial, para o melhor aproveitamento dessas denúncias. Destaca-se que o grande desafio do trabalho está nos textos escritos em linguagem muito informal, contendo muitos erros morfossintáticos, o que dificulta o uso de bibliotecas disponíveis publicamente. Assim, este trabalho conduziu, preliminarmente, uma investigação de pré-processamentos nos textos para maximizar a classificação baseada em *Support Vector Machine* (SVM). A escolha deste algoritmo se deu principalmente pela característica técnica que este algoritmo possui para analisar dados com muitos atributos (*bag-of-words)*, além de já ter sido evidenciado por Aggawar e Zhai (2012), após uma análise com vários classificadores de texto disponíveis, que o classificador SVM possui desempenho acima da média para os vários cenários analisados, sob o mesmo contexto técnico deste trabalho.

Dessa maneira, o maior obstáculo para atingir os objetivos acima é o fato do texto das denúncias serem escritos em português informal e apresenta muitos erros ortográficos, sintáticos e semânticos. A investigação de técnicas de pré-processamento de textos, contidos nesses relatos e delitos de crimes, são essenciais no PLN empregados em denúncias, visando avaliar a melhor adequação de uso das bibliotecas disponíveis, desenvolvidas a partir de *corpus* escrito em português formal, ou com boa qualidade gramatical e ortográfica, quando aplicados a documentos escritos em português coloquial e popular, com erros gramaticais e ortográficos.

Há inúmeros trabalhos correlatos considerando classificação de textos em português [Cesar et al., 2019; Ferraira, 2019; Nascimento, 2019; Andrade, 2015; Rossi, 2015], inclusive trabalhos para português coloquial [Stiilpen, 2016]. No entanto, nenhum tão específico como o trabalho correlato mais próximo ao problema descrito acima, no qual os autores [Pinho et al., 2017] desenvolveram um modelo para classificar denúncias feitas por telefone (*call center*) ao DD utilizando o algoritmo *Weightless Neural Network*. Ressalta-se que no caso desse trabalho correlato, as anotações dessas denúncias são feitas por atendentes treinados, que transcreviam a denúncia de forma padronizada, gerando um texto com qualidade muito superior e livre de erros. Destaca-se que a classificação da denúncia desse trabalho é feita pelo atendente, o que dispensaria a classificação automatizada. Além disso, o algoritmo usado não é o mais recomendado para solução de *bag-of-words* [Aggawar e Zhai, 2012], por isso não foi adotado neste trabalho. Assim, pode-se dizer que para o problema que se deseja solução não há trabalho correlato específico.

Este artigo está dividido em mais quatro seções. A seção dois apresenta os conceitos necessários de Aprendizado de Máquina (AM) utilizados neste trabalho, a terceira seção descreve a metodologia desenvolvida e aplicada no trabalho e no estudo de caso, a seção quarto exibe o estudo de caso, e os resultados encontrados e, finalmente, a última seção apresenta as conclusões e perspectivas de novos trabalhos.

2. Fundamentação Teórica Empregada

2.1. Mineração de Texto

O processo de Mineração de Textos (MT) possui duas abordagens, e uma delas, a análise estatística, que se baseia principalmente no cálculo da frequência dos termos do texto, será a abordada neste trabalho [Jurasfsky and Martin, 2020]. Este processo exige algumas etapas que são clássicas para MT, podendo ser vistas de forma macro como: o pré-processamento, a aplicação de algum algoritmo de AM, seguida da análise dos resultados, que serão detalhados a seguir.

- Pré-Processamento: é executado imediatamente após a coleta dos dados, sendo responsável pela sua preparação (formatação e representação) antes de serem processados pelo modelo escolhido. A seguir são apresentadas as técnicas de pré-processamento utilizadas neste trabalho.
  1. Remoção de acentos: etapa em que é feita a remoção de acentos das palavras.
  2. *Case Folding* (ou Descapitalização): todos os caracteres são convertidos para letras minúsculas ou maiúsculas [Jurasfsky and Martin, 2020].
  3. Remoção de dígitos: remoção de todos os dígitos presentes no texto.
  4. Remoção de pontuação: remoção de caracteres de pontuação presentes no texto.
  5. *"Tokenização"* (ou Atomização): quebra do texto em segmentos menores, normalmente em palavras ou *tokens*. Em caso de palavras combinadas, ou separadas por caracteres como "&" e "-", estas devem ser unidas, formando um único *token* [Carrilho Junior, 2007].
  6. Substituição de acrônimos: é feita utilizando um dicionário pré-definido de acordo com o domínio, havendo tanto acrônimos padrão, como os utilizados em redes sociais ("tb" para também e "vc" para você), além de outros do domínio em questão como por exemplo, "pm" para policial militar. Essa substituição evita que termos com o mesmo valor semântico sejam avaliados como palavras distintas.

7. Remoção de *stopwords*: remoção de termos sem valor semântico, como artigos, preposições, conjunções e pronomes, por exemplo [Jurasfsky and Martin, 2020].
8. Correção ortográfica: A correção ortográfica se propõe a detectar possíveis erros de ortografia, a fim de substituir os termos com erro pela provável grafia correta, visando à melhor qualidade dos textos. Nesse trabalho foi utilizado o algoritmo de correção ortográfica de Norvig (2016), e incluiu-se o algoritmo Filtro de *Bloom* [Bloom, 1970], para acelerar o processamento da correção ortográfica (custosa computacionalmente).
9. *Stemming* (ou stemização): É um processo de transformação de uma palavra na sua versão sem prefixos e sufixos, isto é, mantendo apenas o seu radical. Além dos afixos, também são eliminadas características de gênero e número.

Ao final destas etapas de pré-processamento é necessário que o texto destes documentos seja representado numericamente, que ocorre por meio da conversão textual.

- Conversão do Texto: também chamada de vetorização, é uma das principais etapas de pré-processamento, pois nela ocorre a transformação das palavras em vetores numéricos para que então sejam interpretadas pelos modelos de AM. O modelo *Bag of Words* (em tradução para o português, "saco de palavras") é a representação de documentos de forma numérica mais utilizada para MT e Recuperação da Informação (RI). Como o nome indica o documento é representado apenas pelo conjunto de palavras que o compõe, sem discriminação de ordem. Ele é bastante útil para a montagem da matriz de termos e documentos. Esta matriz pode ser representada por meio da Ponderação TF-IDF, que será utilizada neste trabalho é brevemente apresentada nesta seção. A modalidade TF-IDF tem como objetivo transformar os termos de um documento em vetores de peso. Esta é composta pela fusão entre a frequência do termo (TF, *Term Frequency*) e a frequência inversa do documento (IDF, *Inverse Document Frequency*) [Jurasfsky and Martin, 2020].

$$\mathrm{TF(\,i)} = \text{ ocorrências do termo i/total de termos no documento} \tag{1}$$

$$IDF(i) = \log e^{(total\ \ de\ documentos/_{número\ de\ documentos\ com\ o\ termo\ i})} \tag{2}$$

Desta forma, a ponderação TF-IDF representa uma medida mais ampla, pois ela avalia tanto a frequência de um termo no documento, quanto a sua relação de ocorrência em toda a coleção de documentos (ver equação 3).

$$TFIDF(i) = TF(i) * IDF(i) \tag{3}$$

2.2. Aprendizado de Máquina (AM)

O AM é um subcampo da Inteligência Artificial (IA) responsável por construir modelos que, submetidos a um grande volume de dados, são capazes de aprender, tomar decisões e identificar padrões com o mínimo de interferência humana em um grande volume de dados [Han et al., 2011]. No processo de aprendizado, os dados selecionados são divididos em dois conjuntos: o conjunto de treinamento, usado para ajuste ou aprendizagem dos parâmetros do modelo, e o conjunto de teste, usado para mostrar o desempenho final do modelo. Para medir o desempenho dos modelos, que levam a escolha do melhor modelo, é utilizado o conjunto de validação, constituído por uma parcela dos dados de treinamento e ajudam a evitar *overfitting* e *underfitting*. Assim, para se analisar o desempenho do modelo de forma mais robusta, aplicou-se a técnica de validação cruzada [Han et al.,

2011]. Ao final do processo de aprendizado, o modelo com a parametrização mais adequada aos dados, indicado pela melhor média da(s) métrica(s) considerada(s), é escolhido como o melhor.

2.2.1 Máquinas de Vetores de Suporte

O algoritmo de Máquinas de Vetores de Suporte *ou Support Vector Machines* (SVM), em inglês, é um algoritmo de aprendizado supervisionado que foi desenvolvido por Cortes & Vapnik (1995) consiste em criar classificadores lineares capazes de separar conjuntos de dados através de um hiperplano. Seu objetivo é encontrar um hiperplano separador, por meio de vetores de suporte e margem ótima (função Lagrangeana escrita em função dos parâmetros do hiperplano separador). O SVM pode ser utilizado em dados não linearmente separáveis, a partir do uso de funções núcleo (*kernel functions*) [Smola and Schölkopf, 2002], que projetam os dados em espaços dimensionais de maior ordem de grandeza visando adequar os dados ao hiperplano separador. As funções núcleo mais utilizadas são: linear, polinomial, gaussiana, e sigmoidal, detalhes dessas equações podem ser verificados em [Smola and Schölkopf, 2002].

As métricas de avaliação de desempenho tipicamente utilizadas em modelos de classificação, além da acurácia (divisão dos registros corretamente classificados pelo total de registros avaliados), são apresentadas abaixo, onde l é o número de classes, TP (*true positive*), FP (*false positive*) e FN (*false negative*).

$$Prec\_m = \sum_{i=1}^{l} \frac{TP_i}{TP_i + FP_i} \quad (4) \qquad Rec\_m = \sum_{i=1}^{l} \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$Prec\_M = \frac{\Sigma_{i=1}^{l} \frac{TP_i}{TP_i + FP_i}}{l} \quad (6) \qquad Rec\_M = \frac{\Sigma_{i=1}^{l} \frac{TP_i}{TP_i + FN_i}}{l} \quad (7)$$

$$F1\_\mathrm{m} = \frac{2 * Prec.m * Rec.m}{Prec.m + Rec.m} \quad (8) \qquad F1\_M = \frac{2 * Preci.M * Rec.M}{Prec.M + Rec.M} \quad (9)$$

3. Metodologia

O aplicativo para denunciar via *smartphone* do Disque Denúncia RJ, lançado pelo DD em agosto de 2016, registrou 71.580 denúncias até o mês de fevereiro de 2020 [Disque-Denúncia, 2020]. Uma amostra desta base foi divida em duas partes: 80% para treinamento e 20% para testes. Foram aplicadas técnicas de pré-processamento em cada conjunto de dados gerado, destacando-se que o pré-processamento foi investigado a partir de variações nas etapas de pré-processamento.

A Figura 1 apresenta a metodologia proposta e utilizada no estudo e no pré-processamento das bases de textos. Essa sequência inicial é muito importante e recomendada para estudos com bases textuais, sendo recomendada a utilização dessa sequência, que é considerada básica e bastante comum [Jurasfsky and Martin, 2020]. A sequência utilizada nas etapas de pré-processamento geraram as bases pré-processadas (gerando quatro bases a serem avaliadas), que são empregadas ao estudo. A Base 1 foi gerada a partir da sequencia: remoção de acentos, *Case Folding*, remoção de pronomes oblíquos, dígitos, pontuação, substituição de acrônimos, "Tokenização" e remoção de *stopwords*. A remoção de pronomes oblíquos e placas, embora seja uma prática de remoção de *stopwords*, ela é

apresentada de forma destacada, porque é aplicada sobre os dados não "*tokenizados*", visto que a exclusão de *stopwords* convencionalmente é feita sobre os *tokens*. Essa opção foi apenas para reduzir o volume de palavras da base para a etapa de "Tokenização". A etapa de *stemming* aparece duplicada na figura, porque ela pode ser aplicada em circunstâncias distintas, recebendo uma base que sofreu correção ortográfica ou não. Os resultados obtidos através dessa sequência mostraram-se bastante favorável, gerando quatro bases pré-processadas com resultados expressivos em todas elas. Para as etapas de pré-processamento de remoção de *stopwords*, *stemming* e "*tokenização"* foram utilizados métodos da biblioteca *Natural Language Toolkit* (NLTK), que é um conjunto de bibliotecas do Python para processamento de linguagem natural além das bibliotecas Unidecode e Scikit-Learn. Dessa forma, a metodologia utilizada buscou investigar a influência da correção ortográfica e do uso do processo de *stemming* dado que a base apresenta um corpus muito peculiar e com diferenças significantes em relação a um corpus construído a partir de textos literários, que é usado pelas bibliotecas utilizadas. Em situações com estas, a correção ortográfica (Norvig, 2016), por exemplo, pode trazer mais problemas do que solução, pois ao não encontrar a palavra no seu dicionário, sugere a palavra, do seu dicionário, mais próxima à palavra não identificada no documento, e a substitui, inserindo uma palavra não adequada no contexto do documento.

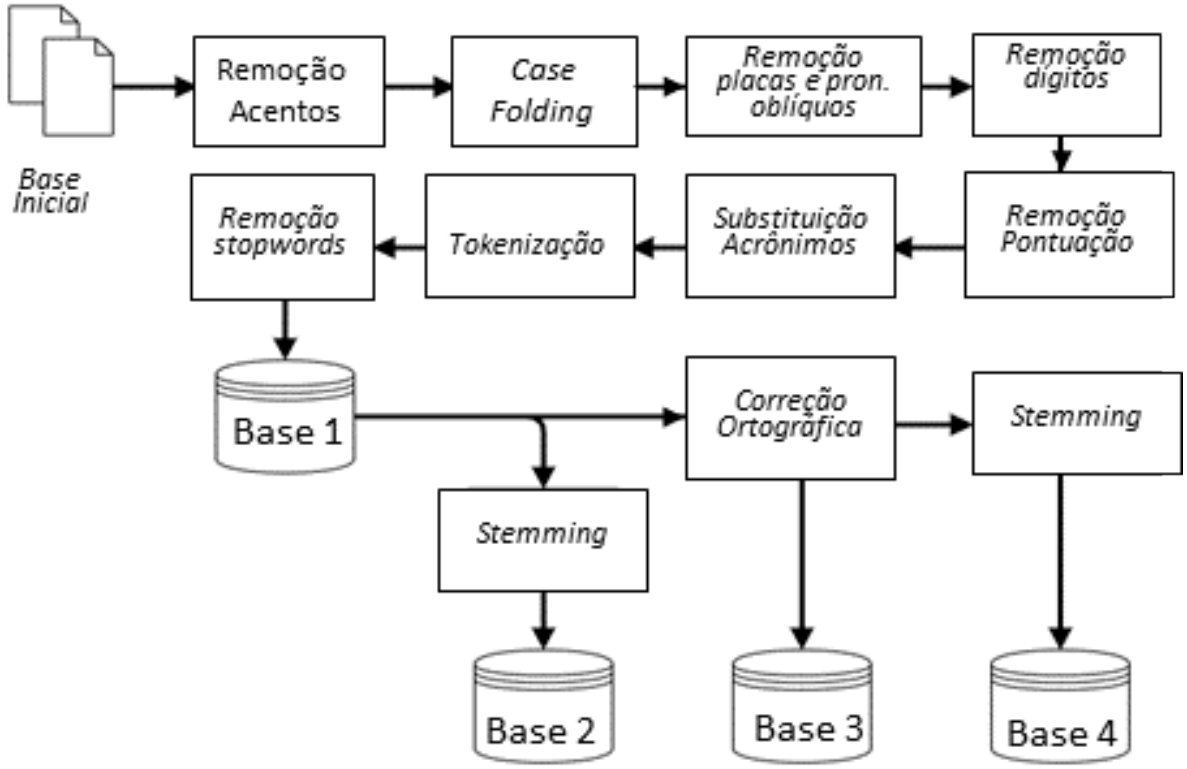

Figura 1: Etapas de pré-processamento e geração de bases pré-processadas.

4. Estudo de Caso

O estudo de caso foi realizado a partir da amostra da base do aplicativo do DD coletada com denúncias realizadas entre os dias 30/07/2016 e 12/09/2017, contabilizando um total de 8.887 denúncias no período, e classificadas em 15 assuntos diferentes: armas; foragidos da justiça, homicídio, homicídios, meio ambiente, outros, procurados; roubo carga/veículo; roubos em geral; terrorismo; tráfico de drogas; tráfico de drogas/armas,

violência contra criança ou adolescente; violência contra mulher ou idoso; violência doméstica. Esses 15 assuntos foram originalmente propostos pelo DD. Durante a validação manual da base do aplicativo, além de correções de documentos mal classificados pelos usuários da ferramenta, foi constatado que alguns tipos de denúncias (ou assuntos) poderiam ser unificados, como os assuntos 'Homicídio' e 'Homicídios'. Outras aglutinações foram necessárias devido ao baixo número de registros e similaridade de contexto, sendo unificadas as classes Violência contra Criança/Adolescente, Mulher ou Idoso na classe Violência Doméstica. Já a classe Terrorismo demonstrou um grande distanciamento do contexto de crimes que ocorrem no Rio de Janeiro, e por isso foi removida.

4.1. Pré-Processamento

Conforme explicitado no esquema presente na Figura 1, foi desenhada como estratégia de pré-processamento a avaliação de quatro combinações das principais técnicas utilizadas. Estas combinações de pré-processamento dão origem a quatro bases indicadas ao longo desta seção como: Base 1, Base 2, Base 3 e Base 4. A Base 1 contém as etapas básicas de pré-processamento, seguindo as seguintes técnicas: (1) Remoção de acentos, (2) *Case Folding*, (3) Remoção de dígitos, (4) Remoção de pontuação, (5) "*Tokenização*", (6) Substituição de acrônimos e (7) Remoção de *stopwords*, que ocorreu em duas etapas: (a) a remoção sobre o texto não "*tokenizado*", para remover placas de veículos e pronomes oblíquos e, (b) sobre o texto "*tokenizado*", após verificação se cada *token* fazia parte da lista de *stopwords* criada. A Base 2 foi criada a partir da Base 1 e adicionada a etapa de *stemming*. A Base 3 foi criada a partir da Base 1 e adicionada a etapa de correção ortográfica. Já a Base 4 foi criada a partir da Base 3. A correção ortográfica utilizada nas bases 3 e 4 utilizou o algoritmo de Norvig (2016), e, devido à baixa qualidade do texto, foi incorporado ao dicionário desse algoritmo (para evitar que) , o *corpus* do projeto CorPop [Pasqualini, 2018], construído a partir de jornais populares do Rio Grande do Sul. Além disso, em função do contexto, algumas palavras foram incluídas, como, por exemplo: milícia, nomes de comunidades, marcas de carro e moto.

4.2. Escolha de Parametrização do Modelo SVM

Cada uma das diferentes funções núcleos (mencionadas na seção 2.2.1), dispõe de parâmetros que devem ser investigados no processo de avaliação da melhor solução. A variável de folga C foi testada no intervalo [0,025, 100] na função linear. O núcleo gaussiano teve 11 variantes, tendo a variável C valores entre 0,01 a 100 e o parâmetro gama verificado no intervalo [0,0001, 50]. O núcleo sigmoidal teve 12 variantes, com C e gamma variando de 1 a 100. Por fim, o núcleo polinomial (*poly*) avaliou 16 variações, com C variando de 1 a 100 e o parâmetro grau do polinômio variando de 1 a 3. Após as quatro bases serem submetidas à validação cruzada, considerando as combinações exaustivas dos parâmetros do SVM acima descritos, a Tabela 1 indica os 10 melhores resultados de acurácia média para as bases do aplicativo.

Tabela 1: Resultado de validação cruzada da base de dados

Modelo	*Acurácia*	*Prec_m(%)*	*Prec_M(%)*	*Rec_m(%)*	*Rec_M(%)*	*F1_m(%)*	*F1_M(%)*	*Base*
Poly 9	***77,60***	***77,60***	***80,19***	***77,60***	***66,27***	***77,60***	***70,99***	***4***
Poly 8	*77,54*	*77,54*	*77,26*	*77,54*	*67,86*	*77,54*	*71,41*	*4*
Poly 13	*77,47*	*77,47*	*80,05*	*77,47*	*66,19*	*77,47*	*70,89*	*4*

Poly 8	*77,35*	*77,35*	*79,02*	*77,35*	*65,93*	*77,35*	*70,44*	*3*
Poly 14	*77,30*	*77,30*	*77,13*	*77,30*	*67,66*	*77,30*	*71,22*	*4*
Linear 4	*77,27*	*77,27*	*79,04*	*77,27*	*66,36*	*77,27*	*70,71*	*4*
Poly 7	*77,27*	*77,27*	*79,04*	*77,27*	*66,36*	*77,27*	*70,71*	*4*
Poly 12	*77,27*	*77,27*	*77,12*	*77,27*	*67,68*	*77,27*	*71,22*	*4*
Poly 8	*77,25*	*77,25*	*78,34*	*78,34*	*67,26*	*77,25*	*71,28*	*2*
Linear 4	*77,19*	*77,19*	*79,97*	*77,19*	*65,11*	*77,19*	*70,02*	*3*

Através da análise dos resultados exibidos acima, pode-se verificar que o classificador SVM, que utiliza a função núcleo polinomial (*poly*), obteve melhor desempenho do que as demais configurações; a função núcleo sigmoidal não esteve entre as 10 melhores posições para nenhuma das bases estudadas. Os resultados de validação da Base 1 também não estavam entre os 10 melhores modelos de forma geral, tendo obtido as métricas mais baixas de validação, embora com uma margem pequena.

4.3. Resultados

O modelo Poly SVM 8 obteve o melhor desempenho de acurácia média de validação para as Bases 2 e 3. A Base 4 teve como melhor modelo o *Poly* SVM 9, sendo o melhor resultado na classificação geral de métricas de validação do aplicativo, e por isso escolhido como solução e aplicado à base teste, conforme resultados apresentados na Tabela 2. As Figuras 2 e 3 indicam a matriz de confusão normalizada para a Base 4-Teste e Nuvem de palavras dessa mesma base.

Tabela 2: Resultado de teste do modelo selecionado Poly SVM 9 para a Base 4

Modelo	Acurácia	Prec_m	Prec_M	Rec_m	Rec_M	F1_m	F1_M
Poly 9	76,11%	76,11%	75,83%	76,11%	65,95%	76,11%	69,64%

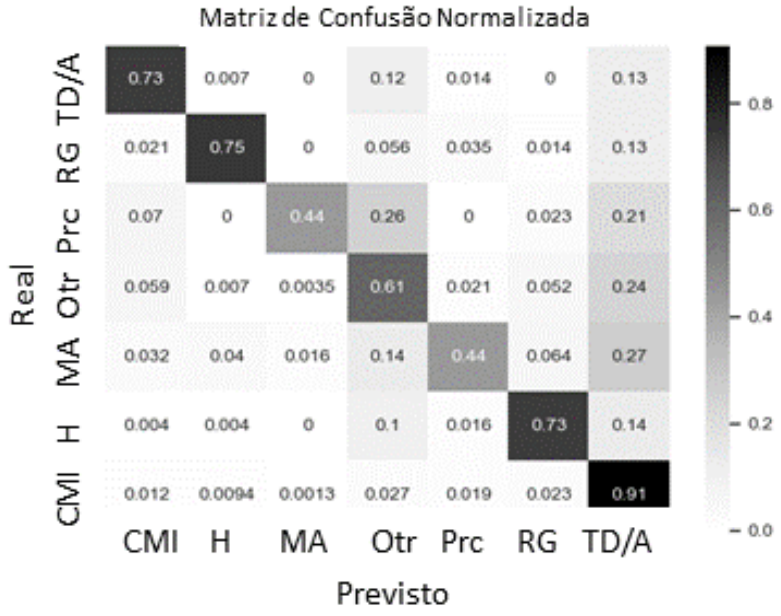

Figura 2: Matriz de Confusão para a Base 4 -Teste

Figura 3: Nuvem de palavras para a Base 4 - Teste

As classes Meio Ambiente (MA), Outros(Otr) e Procurados (Prc) obtiveram, em todas as variações de bases da base, os piores resultados, o que se pode atribuir à diversidade de narrativas que estas classes podem conter. Já a classe Tráfico de Drogas/Armas (TD/A) teve bom resultado em todas as bases, superando as demais classes em percentual de acerto nas classificações, indicando que esta classe tem um vocabulário mais específico e, portanto, favorecendo o modelo de classificação. As classes Homicídios(H), Roubos em Geral (RG) e Crimes contra Crianças, Mulheres e Idosos (CMI) ficaram próximas à acurácia total (76,11%). Como era esperado, na nuvem de palavras (Figura 3) é possível verificar termos ligados à violência e segurança pública.

5. Conclusão e Trabalhos Futuros

Este trabalho teve como objetivo explorar técnicas de PLN em pré-processamento que auxiliassem o desempenho de um modelo de classificação SVM para denúncias criminais do Disque-Denúncia (em bases textuais), visado à automatização na classificação das denúncias, reduzindo o tempo de avaliação manual de cada denúncia. A metodologia consistiu em gerar quatro trilhas distintas de pré-processamento, avaliando para cada uma delas, qual era o conjunto de técnicas de pré-processamento mais adequado, além dos ajustes aos dicionários de dados para superar as dificuldades impostas pelo problema. Os resultados encontrados são positivos e levam a conclusão de que o classificador construído tem boa capacidade de generalização, com métricas satisfatórias, e que as etapas e técnicas de pré-processamento são importantes, devendo ser avaliadas e administradas com cautela. Destaca-se que as bibliotecas NLTK, utilizadas no pré-processamento, são desenvolvidas sobre *corpus* contendo textos formais, baseados na norma culta da língua portuguesa, e, portanto, muito diferentes dos textos observados nas denúncias. Além disso, apesar da base CorPop ter um *corpus* menos formal, trata-se de um grupo pequeno de palavras, para se considerar expressivo no processo de detecção de erros ortográficos, e ainda está distante da forma coloquial e com os muitos erros morfossintáticos dos textos das denúncias do DD.

Como em aplicativos e redes sociais costuma-se reproduzir regionalismos, além de vícios de uma linguagem voltada para a comunicação imediata, longe da norma culta da

língua, está sendo desenvolvido, como extensão desse trabalho, um *corpus* direcionado para a Segurança Pública com base em textos populares e redes sociais, além da utilização de técnicas baseadas em Deep Learning [Zhang et al., 2020] para a produção de vetorização com maior nível de complexidade. Além disso, o modelo aprendido está sendo aplicado às redes sociais, visando identificação de relatos criminais. Essa possibilidade tem por objetivo expandir o número de canais de coleta de informações para a Segurança Pública.

Referências

Aggarwal, C. C., and Zhai, C. X. (2012). A survey of text classification algorithms. In *Mining Text Data* (Vol. 9781461432234, pp. 163-222). Springer US.

Andrade, P.H.M.A. (2015) Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: um Estudo da Automatização da Triagem de Denúncias na CGU, Dissertação em Computação Aplicada da UnB.

Bloom, B.H., (1970). "Space/time trade-offs in hash coding with allowable errors," In: Commun. ACM, vol. 13, no. 7, pp. 422–426.

Carta Capital. (2020) Número de mortos por bala perdida no Rio de Janeiro sobe 23% em 2019. Disponível em: https://www.cartacapital.com.br/sociedade/numero-de-mortos-por-bala-perdida-no-rio-de-janeiro-sobe-23-em-2019/ Acesso em 20-01-20.

Cesar, M. V. G., Vellasco, M. and Figueiredo, K. (2019). Classificação de falhas de equipamentos de unidade de intervenção em construção de poços marítimos por meio de mineração textual. In: XVI Encontro Nacional de Inteligência Artificial e Computacional, 2020, Salvador. Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2019. p. 401-412.

Cortes, C.; Vapnik, V. (1995). Support-vector networks. Machine Learning, v. 20, n. 3, p. 273-297.

Disque-Denúncia (2020) Números. Disponível em: https://disquedenuncia.org.br/o-disque-denuncia /N%C3%BAmeros, Acesso em: 20-02-28.

Ferreira, H.H. (2019) Processamento de Linguagem Natural e Classificação de textos em Sistemas Modulares, Monografia do Departamento de Ciência da Computação da UnB.

Han, J., Kamber, M. and Pei, J. 2011. "Data Mining: Concepts and Techniques" 3rd edition.

Jurasfsky, D.; Martin, J. H. (2008) Speech and Language Processing: An Introduction to Natural Language Processing, Comp. Linguistics, and Speech Recognition. 2st. ed., Prentice Hall USA.

Mendonça, A.V. (2007). Solução de crimes depende de ajuda da população. G1. Rio de Janeiro, p. 00-00. 10 mar. 2007. Disponível em: http://g1.globo.com/Noticias/Rio/0,,MUL9408-5606,00-SOLUCAO+DE+CRIMES+DEPENDE+DE+AJUDA+DA+POPULACAO.html. Acesso em: 12 abril 2021.

Nascimento, R.M.F. (2019).Classificação automática de discursos de ódio em textos do twitter. 2019. 47 f. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Unidade Acadêmica de Serra Talhada, Universidade Federal Rural de Pernambuco, Serra Talhada.

Norvig, P. (2016) How to Write a Spelling Corrector. Disponível em: http://norvig.com/spell-correct.html, Acessado em 2020-06-29.

Pasqualini, B.F. (2018). CorPop: um corpus de referência do português popular escrito do Brasil" Tese Doutorado. Inst. de Letras, Prog. de Pós-grad. em Letras, UFRGS.

Pinho, R., Brito, W., Motta, C. and Lima, P. (2017) Automatic Crime Report Classification through a Weightless Neural Network, European Symp. on Artificial Neural Networks, Comp. Intel. and Mach. Learn., Bruges (Belgium), ISBN 978-287587039-1.

Rossi, R. G. (2015). Classificação automática de textos por meio de aprendizado de máquina baseado em redes. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. doi:10.11606/T.55.2016.tde-05042016-105648.

Smola, A. and Schölkopf, B. (2002) "Learning with Kernels". The MIT Press, Cambridge, MA.

Stiilpen Jr, M. (2016). Um Arcabouço de Processamento de Textos Informais em Português Brasileiro para Aplicações de Mineração de Dados, Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Ouro Preto.

Zhang, A., Lipton, Z., Li, M., and Smola, A.J. (2020). Dive into Deep Learning. https://d2l.ai.