Universidade de Brasília

Faculdade de Tecnologia

Modelagem Automática de Tópicos em Relatos de Violência Doméstica

Bruna Azambuja

TRABALHO DE GRADUAÇÃO

ENGENHARIA DE COMPUTAÇÃO

Brasília

2022

Universidade de Brasília
Faculdade de Tecnologia

Modelagem Automática de Tópicos em Relatos de Violência Doméstica

Bruna Azambuja

Modelagem Automática de Tópicos em Relatos de Violência Doméstica

Trabalho aprovado. Brasília, 28 de junho de 2022:

Prof. Dr. Daniel Guerreiro, UnB/FT/ENE
Orientador

Prof. Dra. Mylène Farias, UnB/FT/ENE
Examinador interno

Prof. Dra. Cristina Castro, UnB
Examinador externo

Brasília
2022

Este trabalho é dedicado a todas as meninas e mulheres que,
em algum ponto, já sofreram preconceito.

“Any sufficiently advanced technology is indistinguishable from magic.”
(Arthur Clarke)

Sumário

Lista de ilustrações

Resumo

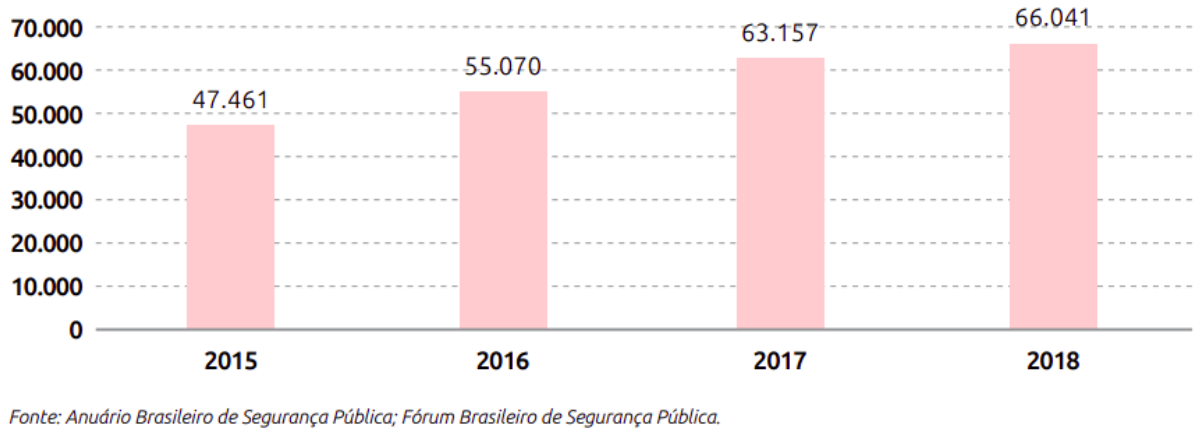
No Brasil, violência doméstica é considerada, desde 1995, uma violação dos Direitos Humanos. Porém mesmo após tal conquista, ainda nos dias de hoje é possível verificar que casos de violência contra a mulher são muito comuns e, em grande parte destes, a mulher não se sente confortável ou segura para reportar tal violência. Faz-se necessário, portanto, um meio anônimo e seguro de capturar os relatos das vítimas sem que haja interferência de terceiros que possa influenciar seu discurso.Neste trabalho, foram analisados os resultados obtidos a partir de conjuntos de dados de violência doméstica usando técnicas de Aprendizado de Máquina para recuperar os assuntos mais recorrentes nos relatos coletados. O Algoritmo de Alocação de Dirichlet Latente - Latent Dirichlet Allocation (LDA) é um dos métodos mais populares para modelagem de tópicos que pode ser aplicado em qualquer área de assunto. O modelo usa uma aproximação Bayesiana para classificar o conjunto de textos em um conjunto de tópicos e pesos. O LDA tem se mostrado um excelente modelo para classificação de tópicos, sendo inclusive utilizado como base para diversos outros algoritmos, e portanto uma ótima opção a ser considerada para uso em bancos de dados que não requerem relacionamentos complexos de tópicos. Após análise detalhada dos resultados obtidos, foi possível verificar que, no contexto de pandemia do COVID-19, os casos de violência familiar, ou seja, violência cometida por parente da vítima, são os casos mais recorrentes. Isto pode ser validado e justificado pelo confinamento da vítima com seu agressor, como será concluído neste trabalho, o que reitera o impacto do algoritmo no contexto mundial de violência, pois os dados podem ser estudados apesar da taxa de notificação de agressões e registros de boletins de ocorrência sejam baixas.

1 Introdução

De acordo com a Organização Mundial da Saúde (OMS), citada por Jaime Alonso em sua publicação *"Gender Violence: A Historical Perspective"* (ALONSO et al., 2014), a violência pode ser definida como o uso intencional da força física ou poder sobre a vítima, em ameaça ou na prática que resulte ou possa resultar em sofrimento, morte, dano psicológico, desenvolvimento prejudicado ou privação. Complementarmente, a Lei 11.340/2006 (LEI. . . , s.d.) determina que a violência doméstica, por sua vez, pode ser caracterizada por qualquer tipo de ação que tenha por motivação o gênero da vítima e que possa causar à mulher morte, lesão, sofrimento físico, sexual ou psicológico, como indicado em estudos correlatos ao tema (SEGURANÇA PÚBLICA, s.d.). Portanto, a violência de gênero tem várias formas de se manifestar: antes mesmo de qualquer contato físico, pode-se presenciar outros tipos de violência, como psicológica, financeira e até casos de perseguição e ameaça.

Esse tipo de violência tem origem cultural e histórica, resultado de uma sociedade patriarcal que desde os primórdios enxerga o sexo feminino como inferior e submisso. A desigualdade entre homens e mulheres, que os colocam em patamares diferentes de poder e respeito, acaba causando a banalização da violência. Até pouco tempo atrás, o que ainda ocorre na atualidade, a lei permitia vários tipos de violência contra a mulher.

Em tempos coloniais no Brasil, era permitido por Lei que o marido assassinasse sua esposa caso este suspeitasse de sua infidelidade. Apenas em 1995 que a violência contra a mulher passou a ser considerada uma violação aos Direitos Humanos no país, conquista derivada da Convenção Interamericana para Prevenir, Punir e Erradicar a Violência contra a Mulher, da qual o Brasil é signatário (CONVENÇÃO. . . , s.d.). Porém mesmo com tal avanço, ainda nos dias de hoje é possível verificar que casos de violência são muito comuns.



Fonte: Anuário Brasileiro de Segurança Pública; Fórum Brasileiro de Segurança Pública.

Figura 1 – Números absolutos de registro de estupro e estupro de vulnerável no Brasil

Como constatado na pesquisa feita pelo *Fórum Brasileiro de Segurança Pública* (SEGURANÇA PÚBLICA, s.d.), representada na Figura 1 , dados registrados de casos de estupro no Brasil vêm crescendo constantemente, alcançando um total de mais de 66 mil pessoas em 2018, sendo 82% desse total do sexo feminino, ou seja, cerca de 54.153 mulheres registraram queixa de estupro apenas naquele ano. Apesar do alto índice, um baixíssimo número de casos de estupro são registrados e notificados à polícia. No Brasil, estima-se, de acordo com Pesquisa Nacional de Vitimização realizada pela Secretaria Nacional de Segurança Pública/Ministério da Justiça em 2013, que cerca de 7,5% das vítimas de violência sexual de fato notificam a polícia. Isso ocorre, dentre outros, pelo medo de vingança por parte do agressor, falta de credibilidade e confiança nas instituições de justiça e até mesmo sentimentos de vergonha e culpa.

Os dados apresentados são com toda certeza preocupantes, porém não é apenas no Brasil que a situação é problemática. De acordo com a UN Woman (UN…, s.d.), organização das Nações Unidas que visa à luta pelo empoderamento feminino e igualdade de gênero, estima-se que, globalmente, uma em cada três mulheres sofreu com algum tipo de violência física ou sexual ao menos uma vez em sua vida, o que gira em torno de 736 milhões de mulheres ao redor do planeta.

Os dados que ilustram a Figura 2, retirados da pesquisa *"Killings of women and girls by their intimate partner or other family members"*, realizada pelo UNODC *(United Nations Office on Drugs and Crimes)* (DRUGS; CRIMES, 2021) evidenciam um cenário mundial mais recente, contabilizando homicídios cometidos apenas por parceiros e familiares, e, ainda, reforça a necessidade de medidas protetivas.

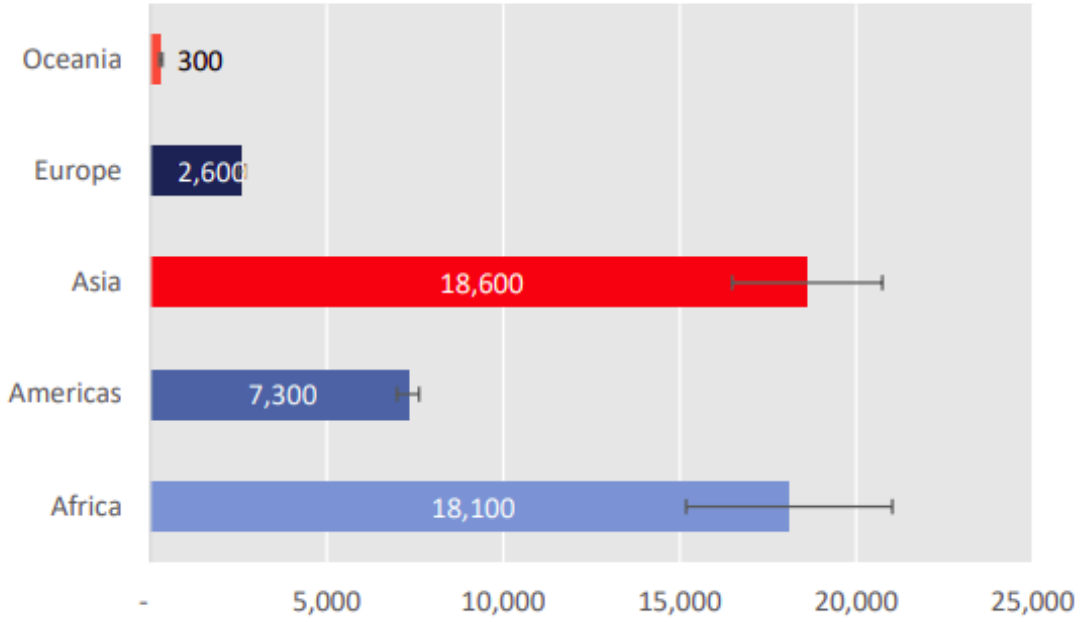


Figura 2 – Número estimado de mulheres vítimas de homicídio por parceiro íntimo/familiar - 2020

1.1 Proposta

Considerando o contexto de violência doméstica agravada no Brasil e no mundo apresentado previamente, este trabalho visa utilizar métodos de Aprendizado de Máquina e Modelagem de Tópicos para estudar e analisar padrões em relatos de vítimas que sofreram este tipo violência.

Para alcançar este objetivo, foi utilizada a base de dados extraída da plataforma *Our Wave*, que será melhor detalhada na seção 1.2. Tal base apresenta relatos produzidos pelas próprias vítimas que sofreram a violência, e portanto não dispõe de nenhum tipo de viés ou interferência de terceiros na informação.

Será utilizado o modelo de Alocação Dirichlet Latente para extrair os tópicos mais presentes na base de dados escolhida, bem como serão aplicadas técnicas de análise de performance como a Pontuação de Coerência, para selecionar, dentre os modelos produzidos, aquele com melhor desempenho.

Na Figura 3 tem-se o Diagrama de Blocos do processo que será aplicado neste trabalho, no qual é demonstrado um exemplo de saída que se é esperado do modelo de Modelagem de Tópicos.

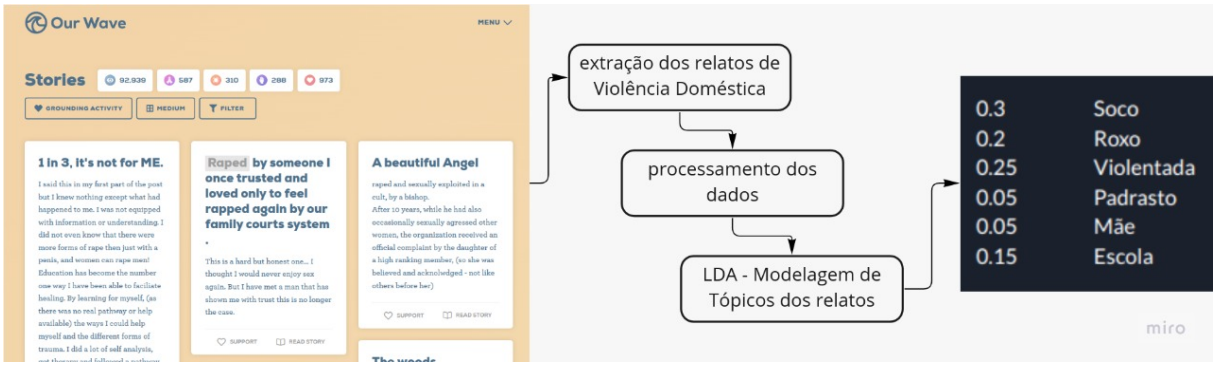


Figura 3 – Diagrama de blocos do Trabalho

1.2 Plataforma *Our Wave*

Um dos problemas dos dados sobre violência doméstica é que a maior parte dos casos nem chegam a ser relatados, como apresentado anteriormente, por diversos motivos como sentimento de vergonha, culpa e falta de credibilidade por parte dos profissionais que recebem os relatos. Numa tentativa de amenizar esse fenômeno, foi criada uma comunidade chamada *Our Wave* (OUR…, s.d.) com o objetivo de criar um ambiente seguro em que as mulheres possam compartilhar suas histórias, suas experiências e relatos de violência de forma completamente anônima, para que possam se sentir ouvidas e acolhidas, e principalmente, que não estão sozinhas.

Our Wave é uma plataforma com o principal objetivo de acolher e amparar mulheres, criando uma comunidade de apoio e compreensão que não só empodere as vítimas de violência, mas também eduque e mobilize a sociedade no âmbito do problema, criando um fórum de discussão delineando possíveis métodos de prevenção e parceria.

Esta plataforma está aberta não só para relatos de violência, mas para todo tipo de manifestação de apoio e suas várias formas de apresentação. Estão disponibilizados uma série de aparatos para que os usuários possam demonstrar apoio individualmente em cada história, assim como recursos selecionados de contatos para organizações similares, que provém uma rede de assistência para escutar e acolher essas vítimas.

O processo de registro dos relatos foi cuidadosamente estruturado para que seja completamente anônimo, portanto a vítima não precisa compartilhar nenhuma informação que a deixe desconfortável - tais como título da história, faixa etária ou categorizar o tipo de agressão ou agressor. Ao submeter sua história, a mulher é livre para escolher se permite que esta seja compartilhada anonimamente em estudos e pesquisas sociais, desta forma, ajudando para que esse tópico tenha a visibilidade necessária em discussões sobre prevenção pública contra a violência doméstica.

1.3 Objetivos

Neste trabalho, busca-se uma análise mais aprofundada da violência doméstica, com o intuito de analisar e classificar, por meio de um classificador externo e imparcial, os relatos de vítimas que sofreram algum tipo de violência, como os citados no início deste capítulo.

Com este classificador desenvolvido, será possível aplicá-lo em bases de dados que contenham relatos produzidos pelas vítimas, para futuramente delinear um padrão com respeito aos assuntos abordados por cada relato. A partir dos resultados, seria possível recuperar informações sobre o comportamento geral de violência para utilizá-las no reforço e na aplicação em políticas públicas de prevenção à violência e proteção à mulher, focalizando os esforços para o contexto de maior frequência produzido pelo modelo.

1.4 Organização do Trabalho

O restante desta monografia se organiza da seguinte forma: o Capítulo 2 provê uma visão global da teoria necessária para o entendimento do trabalho proposto, através de uma breve apresentação sobre Processamento de Linguagem Natural, seguida por uma introdução à Modelagem de Tópicos e ao algoritmo de Alocação de Dirichlet Latente (LDA, do inglês *Latent Dirichlet Allocation*); em seguida, o Capítulo 3 apresenta o passo a passo da solução adotada neste trabalho, com a extração e pré-processamento da base de dados e a implementação do modelo LDA, cujos resultados são apresentados e discutidos no Capítulo 4. Por fim, as conclusões e considerações para trabalhos futuros são elaboradas no Capítulo 5.

2 Fundamentação

Neste capítulo será feita a introdução das bases técnicas necessárias para o estudo presente neste trabalho. Como anteriormente mencionado, para alcançar o objetivo delineado na seção 1.3, foi proposta uma solução com base em análise textual utilizando aprendizado de máquina e modelagem de tópicos. Para tal, é preciso entender a natureza desses conceitos com maior clareza, bem como definir termos técnicos utilizados na descrição do escopo da solução.

2.1 Processamento de Linguagem Natural

Como defende Gallagher (GALLAGHER; RAFFERTY; WU, s.d.), Processamento de Linguagem Natural - ou, em inglês, *Natural Language Processing* (NLP) é a área da computação que permite que sistemas inteligentes entendam a linguagem utilizada por seres humanos e sejam, ainda, capazes de tirar conclusões sobre o que foi dito. Esta linguagem humana não consegue naturalmente ser interpretada por um computador; portanto, uma frase não tem significado real para uma máquina sem que antes seja feito um processamento para que a entenda.

Esta área de pesquisa começou a tomar forma nos anos que se seguiram à II Guerra Mundial, com a necessidade de alguma solução que traduzisse automaticamente o texto de uma linguagem para outra - principalmente de inglês para russo e vice versa, tendo como base o cenário da Guerra (GALLAGHER; RAFFERTY; WU, s.d.). Já na década de 50, Alan Turing, conhecido como o Pai da Computação, publicou o artigo “Computing Machinery and Intelligence” descrevendo um teste, de forma inédita, que deveria ser feito para verificar se uma máquina é de fato inteligente, reforçando mais ainda a importância do estudo de NLP no cenário da época (TURING, s.d.).

O Teste de Turing consiste em um observador externo analisando respostas reproduzidas por um ser humano e por um computador por meio escrito, para tal o computador deve ser capaz de se comunicar com naturalidade; caso o observador não seja capaz de diferenciar a máquina do ser humano baseado em suas respostas, tem-se então, na visão de Turing, uma máquina inteligente. O artigo de Turing apresenta até os dias atuais uma grande relevância no estudo de Aprendizado de Máquina, e principalmente, no estudo de Processamento de Linguagem Natural.

Infelizmente a pesquisa desta área foi gradualmente se tornando inviável devido a falta dos recursos necessários e o limitado poder computacional disponível na época. Em 1966, com a publicação do Relatório ALPAC - *Automatic Language Processing Advisory Committee, 1966* (PIERCE et al., s.d.), que buscou investigar a viabilidade entre o estudo e desenvolvimento de sistemas inteligentes ou recorrer a tradutores humanos, chegou-se à conclusão que tradutores eram a opção mais econômica e, como consequência, o governo dos Estados Unidos diminuiu drasticamente o financiamento para esta área de estudo, resultando em uma gradual suspensão de iniciativas de pesquisa em Processamento de Linguagem Natural (JONES, s.d.).

O cenário melhorou no início dos anos 80 com uma onda de novas ideias que surgiram no estudo de Inteligência Artificial. Abordagens linguísticas com regras complexas do processamento de linguagem natural foram substituídas por abordagens de estatística pura, produzindo novos modelos estatísticos capazes de fazer predições suaves e probabilísticas (GALLAGHER; RAFFERTY; WU, s.d.).

Desde a década de 90 que a área de NLP reconquistou sua importância, devido ao grande fluxo de textos online, e passou a ser foco de pesquisadores ao redor do mundo que começaram a desenvolver soluções de reconhecimento de fala utilizando técnicas de aprendizado de máquina recém elaboradas, como o Aprendizado Profundo - *Deep Learning*. Esta nova técnica consiste em uma ramificação do aprendizado de máquina baseado em sistemas de redes neurais profundas, ou seja, um grafo de várias camadas de processamento aninhadas, capazes de processar dados não estruturados, como textos e imagens, que até então não eram bem representados por técnicas de aprendizado de máquina clássicas (JANIESCH; ZSCHECH; HEINRICH, 2021).

O estudo recente revelou-se uma solução digna de ser explorada na área de linguagem, pois o algoritmo, sem manipulação do programador, deduz o processo de mapeamento de uma entrada para uma saída. Tal comportamento contorna a problemática de ambiguidade das palavras pois não exige que o programador forneça a regra para representar todos os significados possíveis de cada palavra (JOHRI et al., 2021).

Técnicas de NLP como as aqui mencionadas permitem diversas das aplicações mais utilizadas hoje em dia, como Google Tradutor e ferramentas de comunicação como a Alexa, assistente virtual desenvolvida pela Amazon. Tais tipos de tecnologia de computadores pessoais estão por toda parte, incentivando cada vez mais o desenvolvimento desta área de pesquisa que apresenta um constante crescimento (JOHRI et al., 2021). Além das aplicações apresentadas pelas referências citadas, algumas das possíveis análises por meio de Processamento de Linguagem Natural são: extração de sentimento textual para estudo de satisfação de clientela,

pesquisa com preenchimento automático, filtragens de e-mails para classificação de *spam*, e finalmente, a aplicação estudada neste trabalho, extração de tópicos em base textual.

2.2 Modelagem de Tópicos

Uma das grandes áreas dentro de Processamento de Linguagem Natural, a Modelagem de Tópicos, utiliza aprendizado de máquina para ler e absorver um conjunto de documentos, denominado de *corpus*, em que cada documento representa um artigo, um relato ou um livro, por exemplo. O modelo seleciona os tópicos, que podem ser interpretados como assuntos mais discutidos neste conjunto, descobre padrões de uso de palavras e como conectar documentos que compartilham padrões semelhantes. Seu principal objetivo é classificar cada documento em pares (peso-tópico) dos assuntos mais presentes na base disponibilizada, onde cada tópico se apresenta como uma distribuição de probabilidades de palavras (ALGHAMDI; ALFALQI, 2015).

De acordo com (JANIESCH; ZSCHECH; HEINRICH, 2021), modelos podem ser classificados como aprendizado de máquina não supervisionado quando não precisam de uma base de entrada de dados já previamente rotulados, e retornam conhecimentos que o algoritmo aprendeu por si próprio detectando padrões a partir de sua estrutura, representando, portanto, uma solução rápida e prática amplamente utilizada em situações diversas para a extração de assuntos sem necessidade de muito tempo gasto com pré-processamento.

Em contrapartida, podem ser classificados como modelagem de tópicos supervisionada se for necessária uma base de dados rotulada de acordo com os tópicos aos quais o pesquisador quer classificar. Neste caso, o algoritmo retorna classificações para novos dados apenas dentre aqueles tópicos já aprendidos.

Existem inúmeras abordagens para o problema proposto, em que sua performance varia de acordo com a adequação da base de dados às nuâncias do modelo utilizado, tais como tamanho de cada documento e relacionamentos complexos entre as palavras, entre tópicos e ao longo de uma variável temporal (VAYANSKY; KUMAR, 2020).

A procura do modelo ideal para a base de dados selecionada é indispensável para a obtenção de resultados significativos e pequenas mudanças na abordagem utilizada podem acarretar grandes diferenças nas análises produzidas. Tais modelos podem utilizar dados como contagem individual de cada palavra por documento, agrupamento de palavras semelhantes e distância entre palavras para produzir os tópicos mais relevantes, cada qual com sua técnica.

Para a base de dados tratada, que consiste em relatos de tamanho médio e ausência de relacionamentos complexos entre tópicos, a abordagem escolhida foi o modelo probabilístico generativo Alocação de Dirichlet Latente - *Latent Dirichlet Allocation*, um modelo Bayesiano hierárquico que utiliza a Distribuição de Dirichlet como base (VAYANSKY; KUMAR, 2020), e será melhor explicado nas próximas seções.

2.2.1 Distribuição Dirichlet

De acordo com (WILD, 2006), uma Distribuição Estatística pode ser definida como uma lente através da qual vemos a variação nos dados e exploramos a natureza dessa variação. Ou seja, é um modelo matemático que descreve o comportamento de fenômenos aleatórios, cada qual com suas particularidades e regras.

A Distribuição Beta, por exemplo, é uma distribuição aplicada para modelar o comportamento de uma única variável aleatória. Esta variável é mapeada a uma probabilidade entre 0 e 1, de acordo com

$$f(x, \alpha, \beta) = constante \cdot x^{\alpha-1}(x-1)^{\beta-1} \tag{2.1}$$

em que $0 \leq x \leq 1$ é o domínio da variável aleatória, e α e β são parâmetros positivos do modelo, como explicado em (LIN, 2016). A constante representada na equação diz respeito à constante de normalização, que é usada para reduzir qualquer função de probabilidade a uma função de densidade de probabilidade e garante que a probabilidade total seja um.

A Distribuição Dirichlet, por sua vez, é a generalização multivariável da Distribuição Beta. Como pode ser visto em

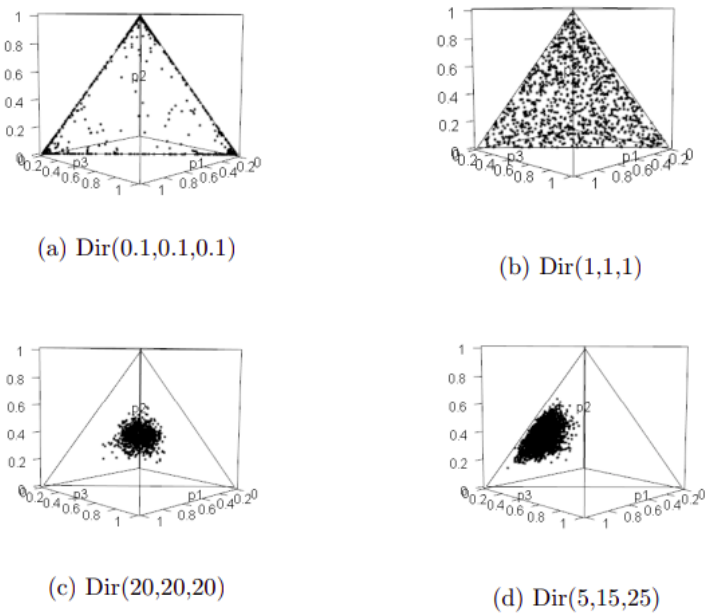
$$Dir(x_1, ..., x_k; \alpha_1, ..., \alpha_k) = \frac{\Gamma(\sum \alpha_k)}{\Pi_k \Gamma(\alpha_k)} \Pi x_k^{\alpha_k - 1} \tag{2.2}$$

estende-se esta ideia para múltiplas variáveis aleatórias, mapeando portanto um vetor de valores aleatórios que são delimitados entre 0 e 1, e cuja soma resulta em 1 (MINKA, 2000). Portanto, em suma, a Distribuição Dirichlet pode ser pensada como uma distribuição de distribuições Beta, conforme pode ser visualizado quando comparamos as Equações 2.1 e 2.2, visto que uma é o produtório sobre um vetor da outra.

De acordo com (LIN, 2016), esta Distribuição pode ser denotada por $Dir(\alpha_k)$, sendo definido por k variáveis aleatórias x_k positivas em relação a um espaço k-dimensional. O parâmetro α determina a esparsidade dos eventos por esse espaço, portanto, é o parâmetro que ditará se os eventos estarão uniformemente distribuídos ou não.

Para elucidar melhor a situação, podemos fazer uso da Figura 4, retirada do artigo *"On*

The Dirichlet Distribution" (LIN, 2016). Na Figura tem-se 1000 pontos plotados num espaço tridimensional gerados pela Distribuição $Dir(\alpha^3)$. Quando todos os α_k são iguais a 1 chamamos de Distribuição Simétrica e sua densidade de distribuição dos k componentes são simetricamente distribuídos pelo espaço k-dimensional - Figura 4 (b). Já no caso em que α_k possuem o mesmo valor e $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$, sua densidade se acumula nas bordas do espaço - Figura 4 (a). Por fim, à medida que o valor de α_k aumenta, sua densidade passa a ser mais concentrada no centro do polígono, podendo variar para a borda de um dos lados caso os valores de α_k não sejam idênticos - Figura 4 (d).



(a) Dir(0.1,0.1,0.1)

(b) Dir(1,1,1)

(c) Dir(20,20,20)

(d) Dir(5,15,25)

Figura 4 – Modelo gráfico Distribuição Dirichlet, retirado de (LIN, 2016)

Esta Distribuição possui várias aplicações conhecidas em diferentes campos de estudo, como no cálculo de probabilidades de correspondência forense de várias populações distintas ou na modelagem de comportamendo de compra de um consumidor (LIN, 2016). Esta modelagem poderia ser equivalente à modelagem de comportamento e padrão de violência, que é um dos objetivos futuros deste trabalho, permitindo desta forma antever tais comportamentos. A Distribuição é usada, dentre outras aplicações, em algoritmos de classificação de tópicos em bases textuais, em especial no modelo de Alocação de Dirichlet Latente, que usa Dirichlet e sua distribuição de probabilidades no espaço para alocar palavras em grupos de tópicos, e tópicos em grupos de documentos, como será melhor especificado a seguir.

2.2.2 Latent Dirichlet Allocation - LDA

O modelo de Alocação de Dirichlet Latente - *Latent Dirichlet Allocation* (LDA) é um algoritmo desenvolvido por David M. Blei, Andrew Y. Ng e Michael I. Jordan no ano de 2003 (BLEI et al., 2003) com proposta de classificação textual por variáveis latentes de tópicos representados por uma distribuição de palavras.

O modelo LDA assume três níveis de hierarquia para a classificação de linguagem natural. Os documentos são o nível mais alto e são compostos por tópicos (nível intermediário), que por sua vez são compostos por palavras, representando o nível mais básico.

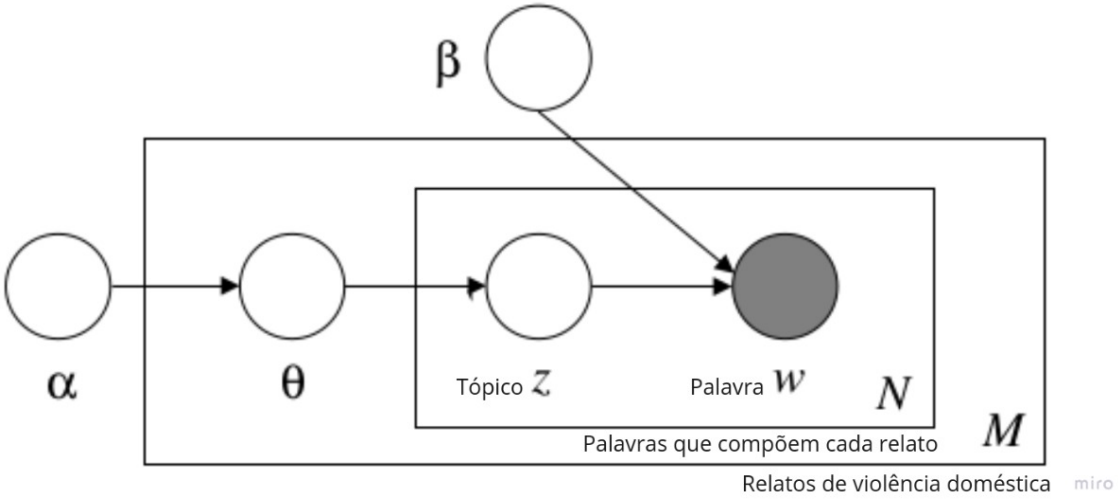


Figura 5 – Modelo gráfico LDA, retirado de (BLEI et al., 2003)

Na Figura 5 retirada do artigo *"Latent Dirichlet Allocation"* (BLEI et al., 2003) pode-se observar o modelo gráfico do LDA. Este tipo de demonstração se chama "Notação de Placa"e é muito utilizada em representações de modelos. Os retângulos na figura representam ações repetitivas no algoritmo e as variáveis no canto inferior de cada um indicam a quantidade de repetições que cada ação será realizada. Os círculos representam variáveis ou parâmetros do próprio modelo: círculos brancos representam variáveis latentes - ou seja, ocultas - e círculos cinzas representam informações dadas pelo problema.

Para o modelo de Alocação de Dirichlet Latente, o retângulo externo M representa a amostragem repetitiva de documentos, já o interno N representa a amostragem repetitiva de tópicos e palavras para cada documento. Utilizando-se do contexto de violência doméstica apresentado neste trabalho, M seriam as histórias relatadas pelas vítimas de violência, e N por sua vez seriam as palavras de cada relato, tais como "roxo", "soco"e "pedi".

Esta demonstração deixa clara a hierarquia tripla do modelo: Documentos, tópicos e palavras. Os parâmetros α e β são hiperparâmetros a nível de corpus e que serão instanciados apenas

uma vez, em que α é um parâmetro para distribuições de tópicos por documento e β é um parâmetro para distribuições de palavras por tópico; Por exemplo: Um β alto indica que cada tópico será uma combinação da maioria das palavras presentes no dicionário, cada qual com seu peso indicando a importância da palavra no tópico correspondente, por sua vez um β baixo indica que cada tópico vai ser uma combinação de só algumas palavras. O mesmo vale para α.

O parâmetro θ_m está a nível de documento, portanto será instanciado a cada documento m presente no corpus; por fim as variáveis z_{mn} e w_{mn} representando cada tópico e palavra, respectivamente, de cada documento, e portanto serão instanciadas $M \cdot N$ vezes, sendo M o número de documentos e N o número de palavras (BLEI et al., 2003).

Definidos em (BLEI et al., 2003), α e β são hiperparâmetros da Distribuição Dirichlet e devem ser selecionados de acordo com a base de dados utilizada para classificação, pois dependem do tamanho do vocabulário e do número de tópicos gerados pelo algoritmo.

Diferentemente de muitos modelos de clusterização para classificação, que apenas são capazes de associar um tópico a cada documento, a representação gráfica do modelo do LDA na Figura 5 deixa claro que é possível amostrar mais de um tópico por documento, e portanto cada documento pode ser associado a vários tópicos ao mesmo tempo, resultando numa classificação mais autêntica e fidedigna visto que um documento usualmente pode tratar de mais de um assunto (VAYANSKY; KUMAR, 2020).

Como abordado em (VAYANSKY; KUMAR, 2020), o algoritmo de Alocação de Dirichlet Latente demonstrou uma melhora significativa nos modelos anteriores pois considera a compreensão de dados não estruturados. Porém uma das grandes desvantagens deste modelo é a necessidade de inferir hiperparâmetros, como o número de tópicos "ideal", considerando principalmente que os resultados finais podem diferir muito a depender das condições iniciais, o que resulta na dificuldade de reproduzir modelos eficientemente.

A solução adotada neste trabalho foi coletar os resultados produzidos iterativamente para várias combinações de hiperparâmetros diferentes, e em seguida analisar os resultados utilizando métricas de qualidade de modelo. Esta solução passa a ser ineficiente à medida que o corpus cresce, porém para este trabalho não apresentou grandes problemas visto que a base de dados a ser estudada era simples e pequena.

A métrica de qualidade utilizada foi a Pontuação de Coerência - *Coherence Score*, que calcula a interpretabilidade do tópico produzido pelo modelo, medindo quão semelhantes são as palavras mais representativas de cada tópico, ou seja, aquelas com maior probabilidade. Esta

métrica é utilizada quando se busca compreensão dos tópicos de modo a serem interpretados por seres humanos (KORENCIC et al., 2021).

De acordo com o artigo *"Exploring the Space of Topic Coherence Measures"* (EXPLORING..., s.d.), para tal, o algoritmo considera as palavras mais importantes de cada tópico e calcula as probabilidades dessas palavras aparecerem juntas na mesma janela de frases, baseado no *corpus* de entrada. Quanto maior essa probabilidade, maior será a Pontuação de Coerência para aquele tópico. Por fim, é feita a agregação dessas pontuações de cada tópico para se obter a pontuação geral do modelo, que pode ser feita a partir de média aritmética dos valores.

O modelo de Alocação de Dirichlet Latente assume a independência entre documentos e palavras, e portanto descarta a relevância da ordem de uma frase e realiza uma análise por cima de relações não complexas entre elas. A suposição de independência entre documentos é inerente à distribuição de Dirichlet e a suposição de independência entre palavras de um documento vem da representação 'Bag of Words', técnica utilizada em muitos modelos estatísticos que consiste em representar um documento como um vetor com a contagem de ocorrências de palavras nele, ressaltando a ideia de desordem entre as palavras (BLEI et al., 2003).

O LDA foi escolhido como método deste trabalho em função de sua capacidade de lidar com documentos grandes e pela dispensabilidade de análise sobre relações complexas entre tópicos.

2.3 Trabalhos relacionados

A ideia de utilizar técnicas de mineração de dados e aprendizado de máquina para análise de tópicos sobre violência doméstica não é inteiramente nova, e conta com alguns artigos já publicados sobre o assunto, que trazem percepções importantes sobre a violência de gênero no mundo, os quais foram utilizados como inspiração para o estudo realizado neste trabalho. É possível analisar o trabalho feito por Jia Xue, Junxiang Chen e Richard Gelles em *"Violence and Gender"* (XUE; CHEN; GELLES, 2019) que trata de tópicos relacionados à violência doméstica em bases de dados do Twitter.

Neste artigo também foi utilizada a técnica de aprendizado não-supervisionado de Alocação de Dirichlet Latente para realizar o estudo sobre os tópicos mais frequentes sendo mencionados na rede social. Conforme relatado em (XUE; CHEN; GELLES, 2019), após os testes realizados foram coletados tópicos mais comuns e frequentes em relatos com a temática de violência doméstica e alguns deles são:

- *violence awareness* - conscientização da violência;
- *greg hardy* - caso de violência doméstica cometida pelo astro da *National Football League* Greg Hardy em 2015;
- *awareness month* - mês da conscientização;
- *victims domestic* - vítimas de violência doméstica;
- *stop domestic* - parar violência doméstica.

Com estes resultados pode-se analisar os tópicos mais discutidos na rede social Twitter a respeito do assunto violência doméstica, assegurando a viabilidade do uso de métodos de modelagem de tópicos, em específico o algoritmo LDA, para mineração de dados de violência de gênero.

Já no intuito de analisar reportagens e notícias sobre violência de gênero durante a pandemia da COVID-19, tem-se o trabalho *"Social mining for sustainable cities"* (MANZOOR et al., 2022), o qual trata justamente sobre as diferenças das reportagens, suas tendências e a natureza da violência de gênero cometidas durante os tempos de confinamento. Este artigo apresenta a mesma proposta deste trabalho, com o objetivo principal de servir como subsídio para ativistas de direitos humanos em sua constante luta contra a desigualdade de gênero e, principalmente, contra a consequente violência gerada por essa desigualdade.

Assim como o anterior, este estudo também foi feito utilizando o algoritmo de Alocação de Dirichlet Latente para extrair os tópicos mais frequentes na base de dados. O resultado obtido, representado na Figura 6, retrata um contexto de violência que chega a ser reportada às autoridades pois apresenta tópicos que abordam casos policiais e julgamentos em tribunal, comportamento já esperado visto que são utilizadas notícias retiradas de jornais sobre violência doméstica. Os tópicos representados nos gráficos da Figura 6 dispõem no eixo Y a contagem de palavras do lado esquerdo, e seu respectivo peso do lado direito do gráfico.

A principal diferença entre os artigos mencionados e o trabalho aqui proposto é a origem da base de dados utilizada. No artigo *"Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter"* (XUE; CHEN; GELLES, 2019) tem-se como base as postagens com o termo-chave "violência doméstica", que são basicamente relatos curtos realizados por terceiros. Em *"Social mining for sustainable cities"* (MANZOOR et al., 2022) tem-se notícias retiradas do jornal sobre ocorrências de violência doméstida durante a pandemia do COVID-19. E, finalmente, no trabalho aqui proposto, serão utilizados como base relatos mais longos e compartilhados pelas próprias vítimas da violência sofrida, por meio da Plataforma Our Wave (OUR..., s.d.).

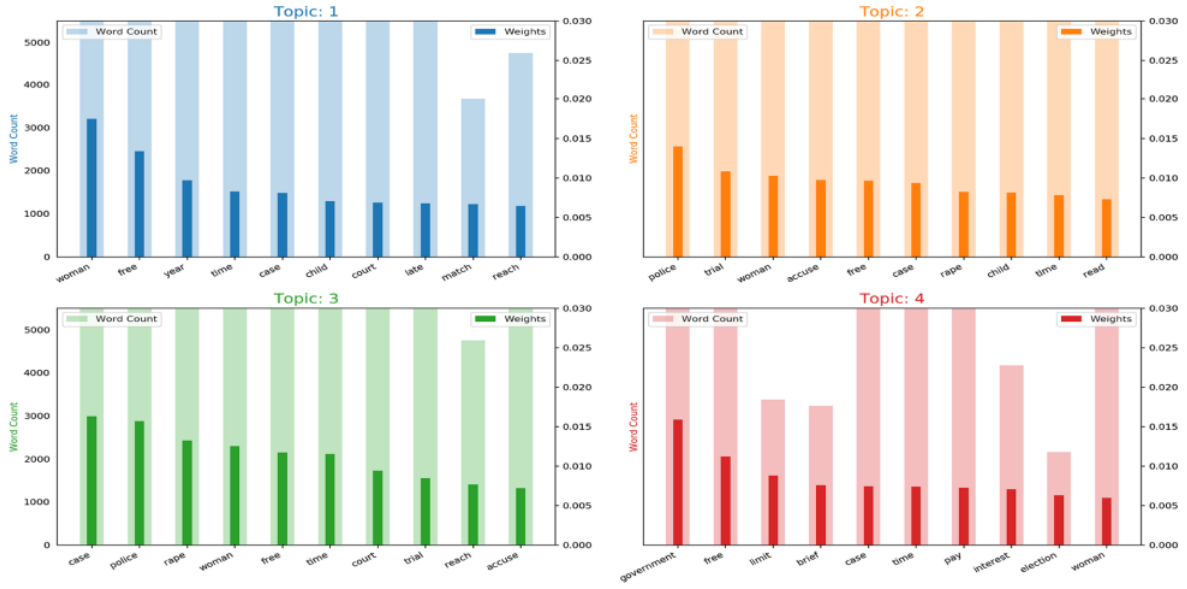


Figura 6 – Tópicos extraídos de Notícias de Violência Doméstica: Adaptado de *Social mining for sustainable cities* (MANZOOR et al., 2022)

Apresentando alguns artigos relacionados, validamos a antingibilidade do objetivo proposto neste trabalho, e partiremos para a metodologia utilizada para alcançar estes objetivos.

3 Metodologia

Neste capítulo será melhor detalhada a metodologia utilizada para a implementação da solução do problema proposto. Como mencionado na seção 2.2, optou-se pela abordagem de modelagem de tópicos utilizando o modelo de Alocação de Dirichlet Latente. Para tal, foi preciso realizar a extração dos relatos da Plataforma Our Wave (OUR..., s.d.) e o pré-processamento destes. A linguagem de programação utilizada foi Python, pela facilidade do tratamento de dados visto que é uma linguagem versátil e com grande suporte à mineração de dados, evidenciado pelas diversas bibliotecas disponíveis com algoritmos já implementados.

A metodologia apresentada será aplicada em duas bases de dados distintas: (a) uma, já estruturada, de notícias, a qual não necessita de nenhum processamento e será utilizada apenas para a realização dos testes iniciais; e (b) base de relatos de violência doméstica já explicitada na seção 1.2, em que a solução será aplicada na íntegra, incluindo a seção de extração e estruturação dos dados ilustrada a seguir.

3.1 Extração dos Dados

A extração dos relatos de violência doméstica da Plataforma Our Wave foi realizada por meio da mineração de dados conhecida por *Web Scraping*, que consiste em criar um processo automatizado que acessa e extrai as informações selecionadas do código HTML da página Web requisitada.

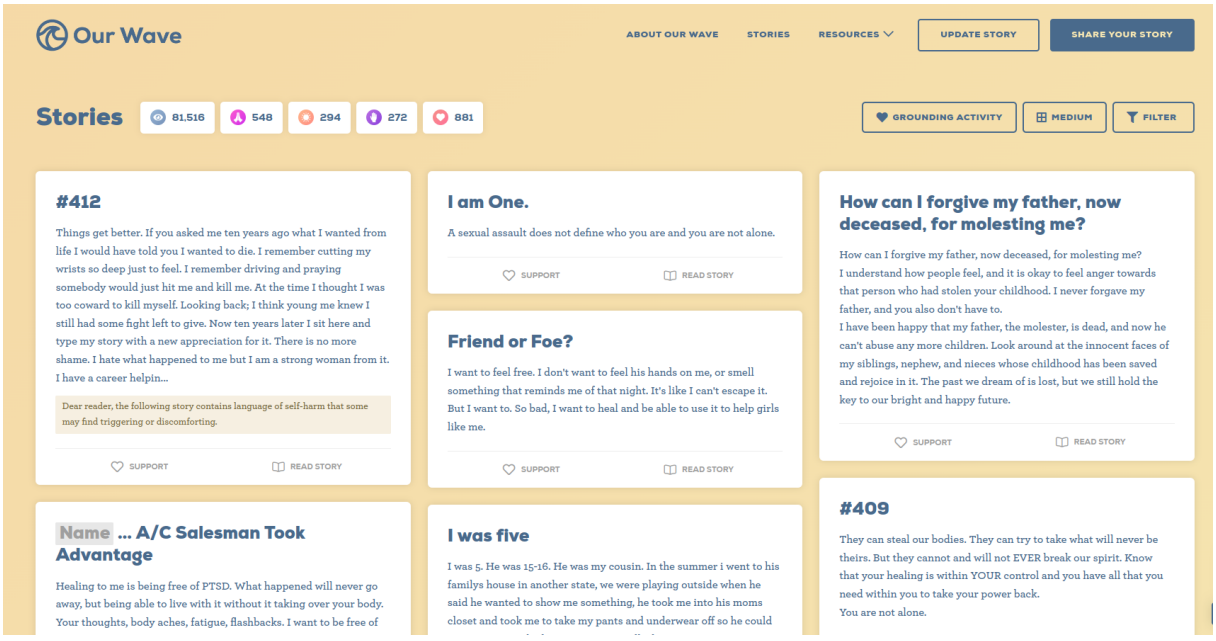


Figura 7 – Disposição da página *Our Wave*: acessada em 12 de Agosto de 2022

A página Our Wave apresenta os relatos das vítimas em uma disposição de cartões em *scroll* infinito, onde cada cartão apresenta uma prévia do relato completo que pode ser acessado clicando no cartão, por meio de hiperlink. O processo pode ser visto nas Figuras 7 e 8. Os relatos são portanto carregados sob demanda, à medida que o usuário navega pela página, por consequência foi necessário simular a navegação até que todos os relatos fossem carregados para, por fim, extrair o HTML completo.

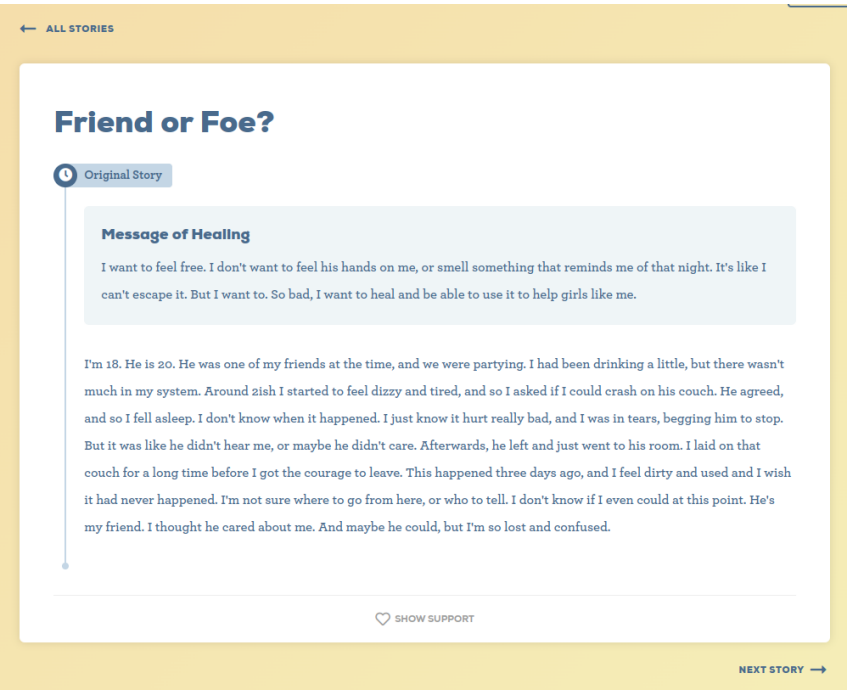


Figura 8 – Exemplo de relato compartilhado na página *Our Wave*: acessada em 12 de Agosto de 2022

Foi utilizada a biblioteca do Python *Beautiful Soup* (RICHARDSON, s.d.) que possui várias funcionalidades já implementadas para acesso de páginas Web e extração de dados. Primeiramente foi preciso instalar o *Web Driver* do navegador Google Chrome, que permite a execução do programa num ambiente do *browser*. Em seguida, utilizando funções do *Beautiful Soup* e do próprio *Web Driver*, foi implementado o *scroll* automático até o fim da página para o carregamento completo de todos os relatos.

Uma vez tendo todos os relatos carregados, inicia-se a extração dos dados. Através de funções do *Beautiful Soup* e por meio de Expressões Regulares, ferramenta da computação que provê uma forma de identificar cadeias de caracteres de interesse, foi possível acessar as páginas individuais de cada relato através de seu respectivo hiperlink, e, em seguida, acessar seu conteúdo. Desta forma, o algoritmo extrai o texto do relato um por um, e grava num arquivo CSV para posterior análise. O arquivo gerado possui os relatos originais das vítimas.

3.2 Pré-Processamento

O pré-processamento consiste em retirar *tokens*, ou seja palavras, que não são de interesse para a classificação dos documentos em tópicos. Para tal, foi utilizada a função *simple preprocess* da biblioteca *gensim* que padroniza os *tokens* convertendo-os para minúsculo e ignorando palavras com menos de dois ou com mais de quinze caracteres.

Em seguida retirou-se as pontuações e palavras vazias, chamadas de *stop words* no meio de Processamento de Linguagem, que são palavras comuns que não possuem relevância significativa na interpretação do assunto, como preposições, pronomes e conjunções. Tal função de remoção de retirada das *stop words* foi importada da biblioteca do Python *Natural Language Toolkit - NLTK* (NATURAL..., s.d.).

Finalmente foi feita a *Lematização* das palavras, processo que consiste em agrupar as formas flexionadas de uma palavra para que possam ser analisadas como uma só. O passo a passo sendo aplicado em um relato original retirado da base de dados pode ser verificado na Figura 9.

Relato Original	"I was waiting at a city bus stop. I saw a man across the street, he was very dirty and obviously not in his right mind. I tried to avoid eye contact but he already saw me. He crossed the street and started shouting at me. Asking for my name and where I lived. I tried to hide in a convenience store until he left. My phone was on 5 percent so I couldn't call anyone to help me. At one point I thought he had left so I walked back out. He was actually hiding behind a sign and came back out once he saw I had come back. I ran back inside the store. Eventually the bus came and I left. Normally I'm not scared of this sort of thing. But because of the state of mind this man was in I didn't know what he would do if I confronted him. When I ran to the bus he started screaming obscenities at me;"
Após Simple Preprocess Gensim	['was', 'waiting', 'at', 'city', 'bus', 'stop', 'saw', 'man', 'across', 'the', 'street', 'he', 'was', 'very', 'dirty', 'and', 'obviously', 'not', 'in', 'his', 'right', 'mind', 'tried', 'to', 'avoid', 'eye', 'contact', 'but', 'he', 'already', 'saw', 'me', 'he', 'crossed', 'the', 'street', 'and', 'started', 'shouting', 'at', 'me', 'asking', 'for', 'my', 'name', 'and', 'where', 'lived', 'tried', 'to', 'hide', 'in', 'convenience', 'store', 'until', 'he', 'left', 'my', 'phone', 'was', 'on', 'percent', 'so', 'couldn', 'call', 'anyone', 'to', 'help', 'me', 'at', 'one', 'point', 'thought', 'he', 'had', 'left', 'so', 'walked', 'back', 'out', 'he', 'was', 'actually', 'hiding', 'behind', 'sign', 'and', 'came', 'back', 'out', 'once', 'he', 'saw', 'had', 'come', 'back', 'ran', 'back', 'inside', 'the', 'store', 'eventually', 'the', 'bus', 'came', 'and', 'left', 'normally', 'not', 'scared', 'of', 'this', 'sort', 'of', 'thing', 'but', 'because', 'of', 'the', 'state', 'of', 'mind', 'this', 'man', 'was', 'in', 'didn', 'know', 'what', 'he', 'would', 'do', 'if', 'confronted', 'him', 'when', 'ran', 'to', 'the', 'bus', 'he', 'started', 'screaming', 'obscenities', 'at', 'me']
Após Remoção de pontuação e Stop Words	['waiting', 'city', 'stop', 'across', 'street', 'dirty', 'obviously', 'right', 'mind', 'tried', 'avoid', 'contact', 'already', 'crossed', 'street', 'started', 'shouting', 'asking', 'name', 'lived', 'tried', 'hide', 'convenience', 'store', 'left', 'phone', 'percent', 'call', 'anyone', 'help', 'point', 'thought', 'left', 'walked', 'back', 'actually', 'hiding', 'behind', 'sign', 'came', 'back', 'come', 'back', 'back', 'inside', 'store', 'eventually', 'came', 'left', 'normally', 'scared', 'sort', 'thing', 'state', 'mind', 'know', 'would', 'confronted', 'started', 'screaming', 'obscenities']
Após Lematização	['wait', 'citi', 'stop', 'across', 'street', 'dirti', 'obvious', 'right', 'mind', 'tri', 'avoid', 'contact', 'alreadi', 'cross', 'street', 'start', 'shout', 'ask', 'name', 'live', 'tri', 'hide', 'conveni', 'store', 'leav', 'phone', 'percent', 'call', 'anyon', 'help', 'point', 'think', 'leav', 'walk', 'back', 'actual', 'hide', 'behind', 'sign', 'come', 'back', 'come', 'back', 'back', 'insid', 'store', 'eventu', 'come', 'leav', 'normal', 'scar', 'sort', 'thing', 'state', 'mind', 'know', 'would', 'confront', 'start', 'scream', 'obscen']

Figura 9 – Exemplo de Pré Processamento

Em seguida, após realizada a retirada de *tokens* indesejados, é preciso alterar a representação do texto de forma a ser interpretada pelo modelo. Como mencionado na seção 2.1, o computador não é capaz de compreender a linguagem natural utilizada pelo ser humano, portanto, antes de passar os dados coletados para o algoritmo, é feita a transformação para um formato compreensível pelo computador. Existem diversas representações conhecidas no contexto de Processamento de Linguagem Natural, em particular a *Bag of Words*, representação escolhida neste trabalho, comum por ser simples e flexível.

O artigo *"An Overview of Bag of Words"* (QADER; AMEEN; AHMED, 2019) explicita a importância desta representação no âmbito de categorização de textos pela simplicidade computacional e conceitual do modelo, em comparação com outros métodos de representação. O método *Bag of Words* consiste em gerar um vetor com o número de ocorrências das palavras de cada documento, desconsiderando a ordem em que aparecem. Tal vetor de frequências pode ser chamado de histograma do documento e está exemplificado na Figura 10.

Documento representado por Bag of Words	[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 4), (7, 1), (8, 1), (9, 1), (10, 3), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1), (16, 1), (17, 1), (18, 2), (19, 1), (20, 1), (21, 3), (22, 1), (23, 2), (24, 1), (25, 1), (26, 1), (27, 1), (28, 1), (29, 1), (30, 1), (31, 1), (32, 1), (33, 1), (34, 1), (35, 1), (36, 1), (37, 2), (38, 1), (39, 1), (40, 2), (41, 2), (42, 1), (43, 1), (44, 2), (45, 1), (46, 1), (47, 1)]	
Documento representado por Bag of Words interpretado pelo Dicionário	Palavra 0 ("across") aparece 1 vez(es). Palavra 1 ("actual") aparece 1 vez(es). Palavra 2 ("alreadi") aparece 1 vez(es). Palavra 3 ("anyon") aparece 1 vez(es). Palavra 4 ("ask") aparece 1 vez(es). Palavra 5 ("avoid") aparece 1 vez(es). Palavra 6 ("back") aparece 4 vez(es). Palavra 7 ("behind") aparece 1 vez(es). Palavra 8 ("call") aparece 1 vez(es). Palavra 9 ("citi") aparece 1 vez(es). Palavra 10 ("come") aparece 3 vez(es). Palavra 11 ("confront") aparece 1 vez(es). Palavra 12 ("contact") aparece 1 vez(es). Palavra 13 ("conveni") aparece 1 vez(es). Palavra 14 ("cross") aparece 1 vez(es). Palavra 15 ("dirti") aparece 1 vez(es). Palavra 16 ("eventu") aparece 1 vez(es). Palavra 17 ("help") aparece 1 vez(es). Palavra 18 ("hide") aparece 2 vez(es). Palavra 19 ("insid") aparece 1 vez(es). Palavra 20 ("know") aparece 1 vez(es). Palavra 21 ("leav") aparece 3 vez(es). Palavra 22 ("live") aparece 1 vez(es). Palavra 23 ("mind") aparece 2 vez(es).	Palavra 24 ("name") aparece 1 vez(es). Palavra 25 ("normal") aparece 1 vez(es). Palavra 26 ("obscen") aparece 1 vez(es). Palavra 27 ("obvious") aparece 1 vez(es). Palavra 28 ("percent") aparece 1 vez(es). Palavra 29 ("phone") aparece 1 vez(es). Palavra 30 ("point") aparece 1 vez(es). Palavra 31 ("right") aparece 1 vez(es). Palavra 32 ("scar") aparece 1 vez(es). Palavra 33 ("scream") aparece 1 vez(es). Palavra 34 ("shout") aparece 1 vez(es). Palavra 35 ("sign") aparece 1 vez(es). Palavra 36 ("sort") aparece 1 vez(es). Palavra 37 ("start") aparece 2 vez(es). Palavra 38 ("state") aparece 1 vez(es). Palavra 39 ("stop") aparece 1 vez(es). Palavra 40 ("store") aparece 2 vez(es). Palavra 41 ("street") aparece 2 vez(es). Palavra 42 ("thing") aparece 1 vez(es). Palavra 43 ("think") aparece 1 vez(es). Palavra 44 ("tri") aparece 2 vez(es). Palavra 45 ("wait") aparece 1 vez(es). Palavra 46 ("walk") aparece 1 vez(es). Palavra 47 ("would") aparece 1 vez(es).

Figura 10 – Exemplo de Documento representado na forma de Bag of Words

O último passo do pré-processamento é a criação de um dicionário filtrando os *tokens* de frequência nos extremos, ou seja, removendo as palavras que aparecem em mais do que uma certa porcentagem dos documentos, que chamaremos de x_{above}, ou em menos de uma certa porcentagem dos documentos, que chamaremos de y_{below}. Os hiperparâmetros x_{above}

e y_{below} foram inferidos empiricamente, da mesma forma que os hiperparâmetros do LDA, numa sequência de iterações com combinações diversas de valores e posterior análise da performance do modelo.

Tal filtragem ocorre pois as palavras presentes em uma grande quantidade de documentos ou em uma pequena porcentagem do *corpus* não retém informações relevante sobre um documento específico, e portanto, podem ser descartadas. Finalizando assim o pré-processamento dos dados que serão utilizados no modelo de classificação de tópicos.

3.3 Implementação do LDA

Quanto ao Modelo de Alocação de Dirichlet Latente foi utilizada a implementação da biblioteca *gensim* (GENSIM..., s.d.), a qual recebe como parâmetro o *corpus* pré-processado e o dicionário filtrado como explicado na seção anterior, o número de tópicos que o modelo deve classificar, e o número de passadas que o algoritmo deve realizar no *corpus*.

Primeiramente dividiu-se a base de dados aleatoriamente em Conjunto de Teste e Conjunto de Treino numa proporção de 20% e 80% do total de documentos na base. Foram implementados laços de repetição para cada hiperparâmetro que deveria ser inferido, incluindo os parâmetros de filtragem do dicionário x_{above} e y_{below} mencionados na seção 3.2, e o número de passadas do algoritmo pelo corpus, ou seja, o número de revisitas que o modelo faz à base de dados durante o treinamento. Para cada iteração treinou-se um modelo utilizando apenas o Conjunto de Treino com uma combinação de parâmetros diferente.

Para cada modelo treinado calculou-se a Pontuação de Coerência - *Coherence Score* mencionada na seção 2.2.1, utilizando a implementação *Coherence Model* da biblioteca *gensim*, e finalmente salvou-se o modelo em diretório local e seus respectivos resultados em arquivo CSV, seguindo a sequência: identificador do modelo, tópicos gerados pelo modelo e Pontuação de Coerência.

A lógica de treinamento foi reproduzida graficamente na Figura 11, em que cada seta representa um laço de repetição e sua variável de controle e, internamente, tem-se os passos de cada etapa do treinamento.

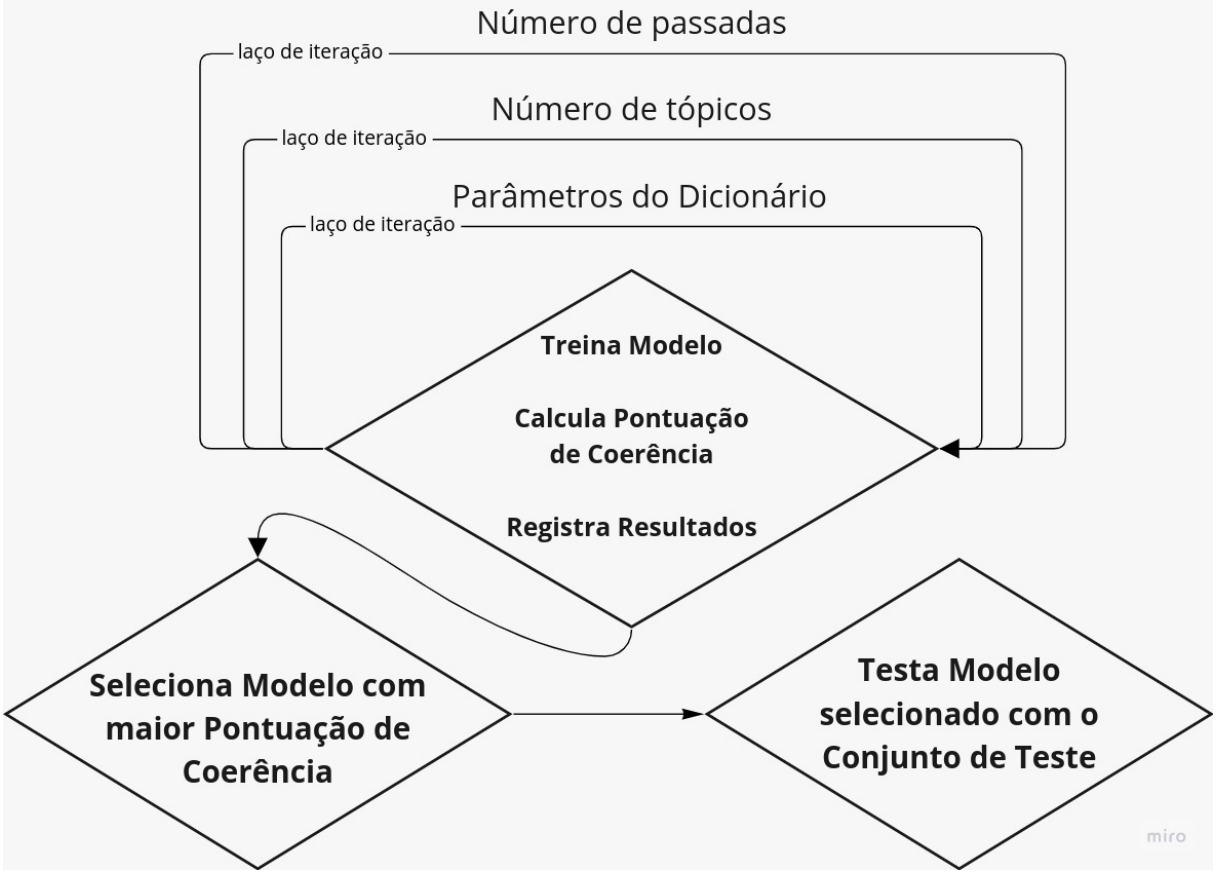


Figura 11 – Representação gráfica do treinamento do Modelo

Após o treinamento, obteve-se 150 modelos treinados com diversas combinações de hiperparâmetros e Pontuações calculadas. Com o arquivo de resultados foi possível realizar o processamento dos dados e extrair o modelo que obteve a maior Pontuação, bem como os tópicos produzidos. Com o modelo que apresentou melhor desempenho, utilizou-se o Conjunto de Teste para realizar a classificação de textos inteiramente desconhecidos pelo modelo, e assim analisar sua verdadeira performance em novos dados.

Os resultados serão apresentados no próximo capítulo.

4 Resultados

Neste capítulo será descrita a análise dos resultados obtidos após o processamento relatado na seção 3.2, para tal serão apresentados os resultados do algoritmo aplicado em uma base de dados de notícias em português, que foi utilizada como teste inicial para a implementação da primeira versão do algoritmo implementada. Em seguida será especificada a adaptação necessária na resolução do problema para a nova base de dados de relatos de violência doméstica, retirados conforme explicado na seção 3.1 e seus resultados.

4.1 Teste Inicial em Base de Notícias

Inicialmente foi utilizada a base de dados pública *"News of the Brazilian Newspaper"* (NEWS..., s.d.) de notícias em português da Folha de São Paulo, entre o período de Janeiro de 2015 e Setembro de 2017. Tal base foi extraída do *Kaggle*, uma comunidade de cientistas de dados que disponibilizam diversos recursos de *data science* e *machine learning*, permitindo que entusiastas publiquem conjunto de dados públicos para treinamento de modelos, usem notebooks integrados à GPU e concorram com outros cientistas para resolver desafios de ciência de dados.

O experimento realizado na base de dados de notícias serviu como teste inicial para validação do processo de treinamento e extração de tópicos de um modelo LDA, e portanto, não contou com a extração de dados explicitado na seção 3.1, bem como a separação em Conjunto de Treino e Conjunto de Teste para posterior validação do modelo, visto que tais análises seriam feitas apenas para a base de dados de violência doméstica.

A base utilizada, *"News of the Brazilian Newspaper"*, possui pouco mais de 167 mil documentos organizados em colunas de arquivo CSV: Título, Corpo da Notícia, Data da Notícia, Categoria, Subcategoria e o link de referência da notícia. Para a classificação feita neste trabalho apenas foi necessário retirar o corpo da notícia para o treinamento e classificação do modelo de Alocação de Dirichlet Latente.

Após a importação da base de dados e extração da coluna que nos interessa, foi feito o pré-processamento descrito na seção 3.2 e seu resultado foi salvo em novo arquivo para posterior treinamento. Obteve-se um dicionário com 268.661 palavras inicialmente e 41.157 palavras após sua filtragem de extremos com os parâmetros x_{above} de 50% e y_{below} de 5% do total de documentos, explicados na seção 3.2.

Utilizando o modelo LDA implementado pela biblioteca *gensim* (GENSIM…, s.d.), foi possível extrair resultados para um conjunto pré-fixado de hiperparâmetros, apenas para efeito de teste, e então foi analisada manualmente a performance do modelo. Os hiperparâmetros foram escolhidos com base no número total de registros na base de dados, sem estudo feito de antemão e apenas visando gerar um conjunto de resultados inicial para familiarização do formato reproduzido pelo LDA.

Os resultados podem ser analisados nas Figuras 12, 13, 14, 15 e 16. Estes resultados foram produzidos por um modelo treinado com 4 passadas, 5 tópicos e parâmetros de filtragem do dicionário explicados na seção 3.2 de 5% do total para y_{below} e 50% do total para x_{above}. No eixo X tem-se as palavras mais representativas de cada tópico e o eixo Y apresenta os pesos das respectivas palavras, ou seja, a medida de relevância da palavra no tópico representado.

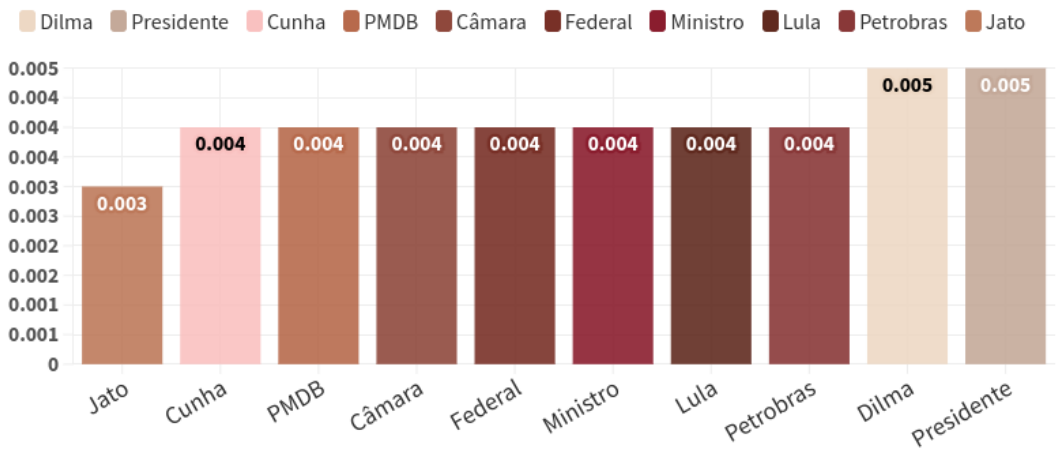


Figura 12 – Tópico 1: Política

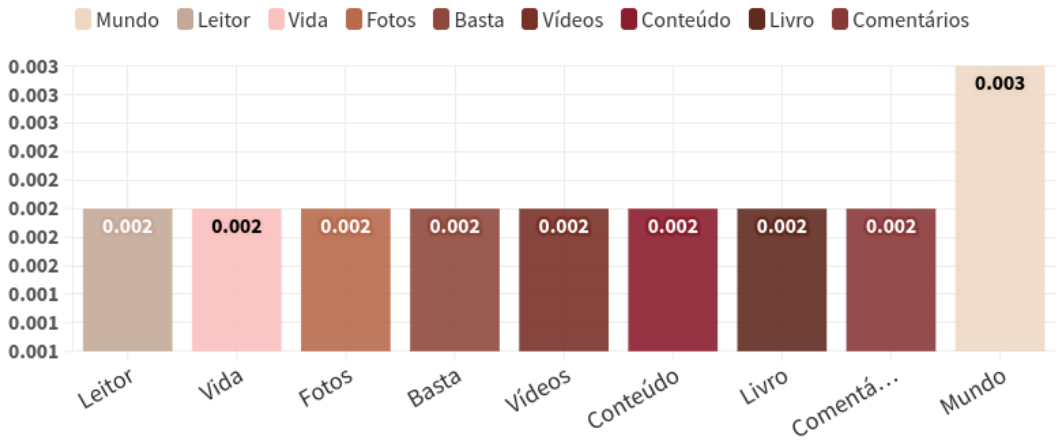


Figura 13 – Tópico 2: Lazer/Arte

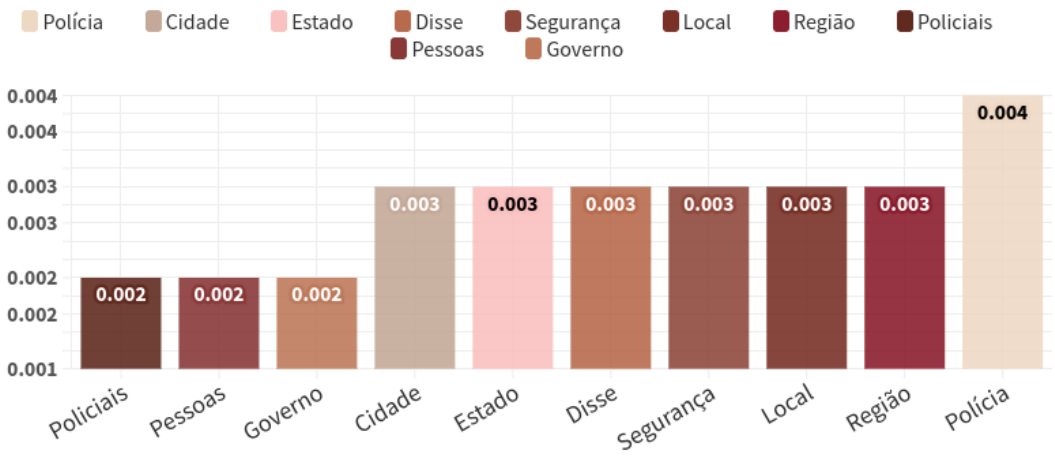


Figura 14 – Tópico 3: Segurança Púlica

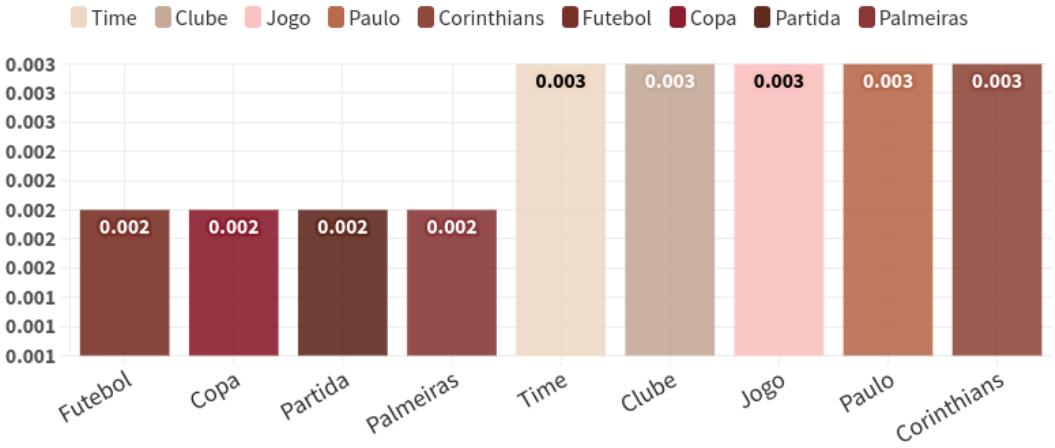


Figura 15 – Tópico 4: Esporte

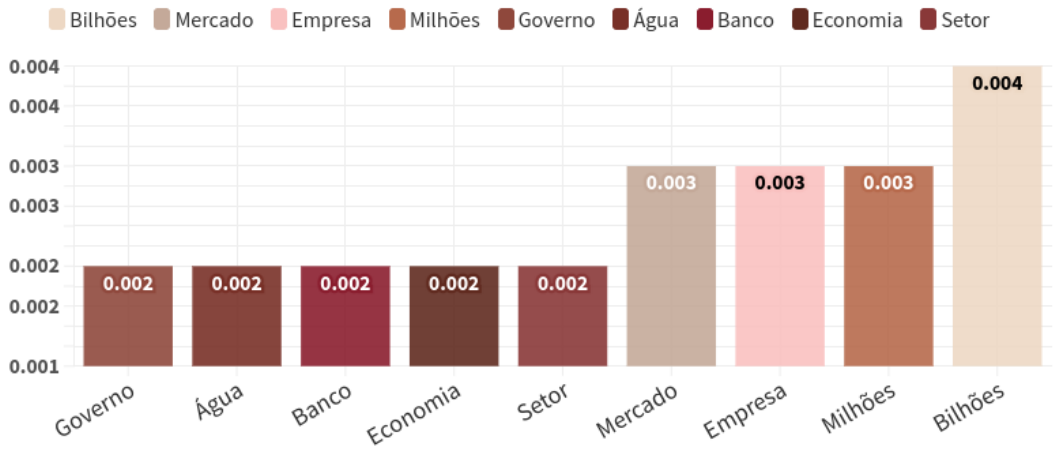


Figura 16 – Tópico 5: Economia

Analisando as imagens dos resultados obtidos, em que cada uma apresenta um tópico retornado pelo modelo com seus respectivos pares peso-palavra, pode-se observar que apesar do modelo não retornar um assunto diretamente e sim um conjunto de palavras, é possível atribuir uma área ao assunto abordado por essas palavras.

Por exemplo, o Tópico 1 representado na Figura 12 aborda principalmente o assunto Política e Governo. Já na Figura 13 fica explícita a conexão das palavras retornadas com o assunto de Lazer e Arte. O mesmo pode ser feito para o Tópico 3 na Figura 14 que aborda principalmente fatores de Segurança Pública. Por sua vez, o Tópico 4 representado na Figura 15 aborda bastante Esporte e Futebol; e finalmente o Tópico 5 na Figura 16 trata da Economia do País.

A distinção clara entre os temas produzidos, mesmo sem a fase referente à experimentação de combinações de hiperparâmetros para melhor explorar a performance do modelo, possivelmente se deve à vastidão da base de dados utilizada, que conta com mais de 167 mil documentos de temas variados, ressaltando o fato que variedade e quantidade são fatores essenciais no treinamento de modelos consistentes.

4.2 Resultados em Base de Violência Doméstica

Após os testes de validação realizados na seção anterior, iniciou-se o real objetivo deste trabalho, utilizando a base de dados de violência doméstica da plataforma Our Wave extraída por meio do processo especificado na seção 3.1. Foram necessários alguns ajustes no código utilizado no teste inicial, visto que a base conta com dados não estruturados e em inglês.

Com apenas 272 registros, o conjunto de relatos de violência doméstica se mostrou menos promissor quanto aos resultados obtidos em relação à base de notícias. Foi então feito o estudo de performance do modelo explicitado na seção 3.3, que utiliza combinações variadas de hiperparâmetros e o cálculo de sua performance através da Pontuação de Coerência, para obter o melhor modelo dentro dos limites proporcionados pela base de dados.

Após dividir a base aleatoriamente em Conjunto de Teste e Conjunto de Treino na proporção de aproximadamente 20% e 80%, resultando em 50 e 222 documentos respectivamente, utilizou-se os seguintes valores de hiperparâmetros:

- Para a quantidade de passadas/visitas pelo Conjunto de Treino para atualização dos pesos foram atribuídos os valores 2, 3, 4, 5 e 6;
- Para a quantidade de tópicos produzidos pelo modelo foram atribuídos os valores 5, 6, 7, 8, 9 e 10;

- Para o parâmetro y_{below} foram atribuídos os valores 1%, 5%, 10%, 15% e 20% do Conjunto de Treino e para o parâmetro x_{above} 50% do Conjunto de Treino, ambos referentes à filtragem de extremos do dicionário explicados em 3.2.

Ao final de todas as combinações possíveis dos hiperparâmetros listados acima, obteve-se um total de 150 modelos salvos em ambiente local, juntamente com suas Pontuações de Coerência, para posterior análise de performance.

Em segundo momento, descartou-se todos os modelos exceto aquele com a maior Pontuação, e obteve-se o valor ótimo de cada hiperparâmetro: 6 tópicos, 6 passadas e 15% do Conjunto de Treino para y_{below}. Obteve-se um dicionário com 171 palavras para tais parâmetros, podendo apresentar uma pequena variância a depender do Conjunto de Treino selecionado aleatoriamente no início do treinamento.

O modelo selecionado produziu uma Pontuação de Coerência de 0,32 e os tópicos a seguir, representados pelas Figuras 17, 18, 19, 20, 21 e 22.

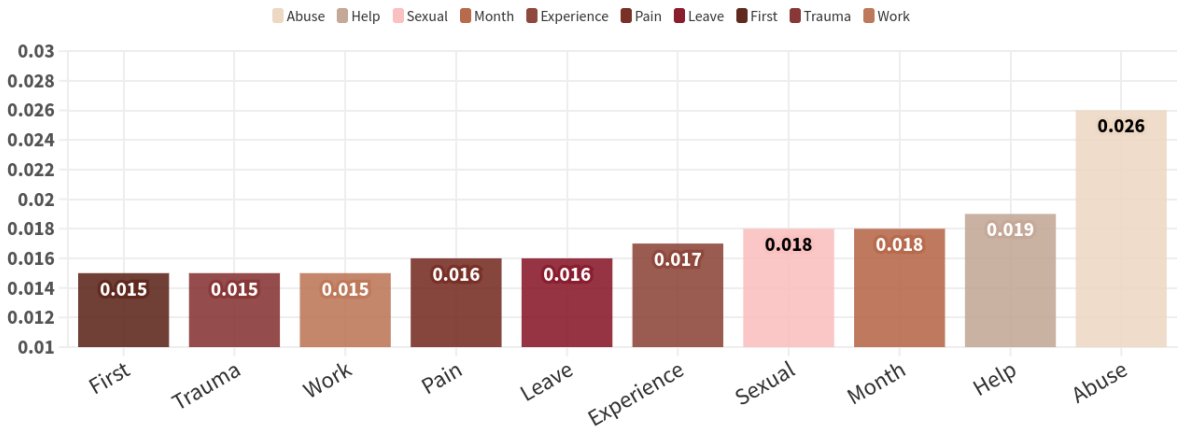


Figura 17 – Tópico 1

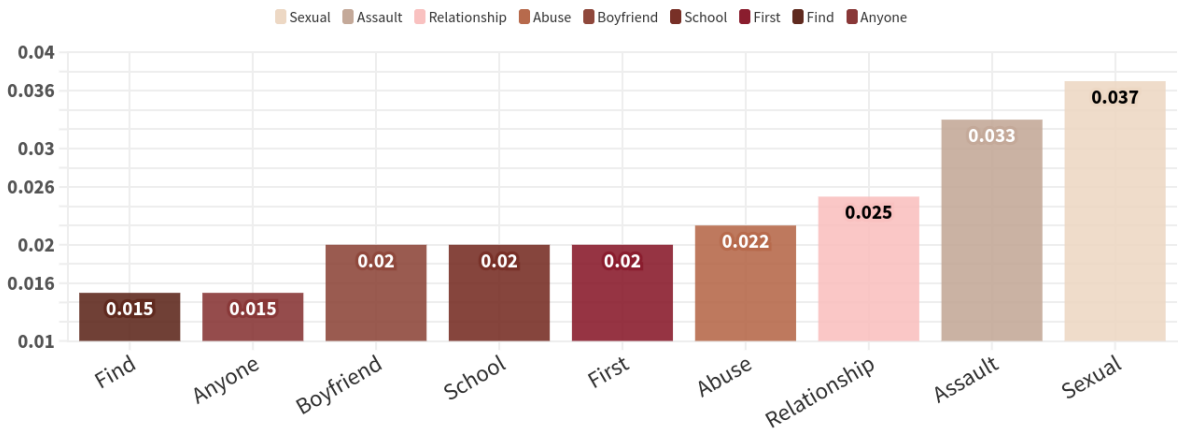


Figura 18 – Tópico 2

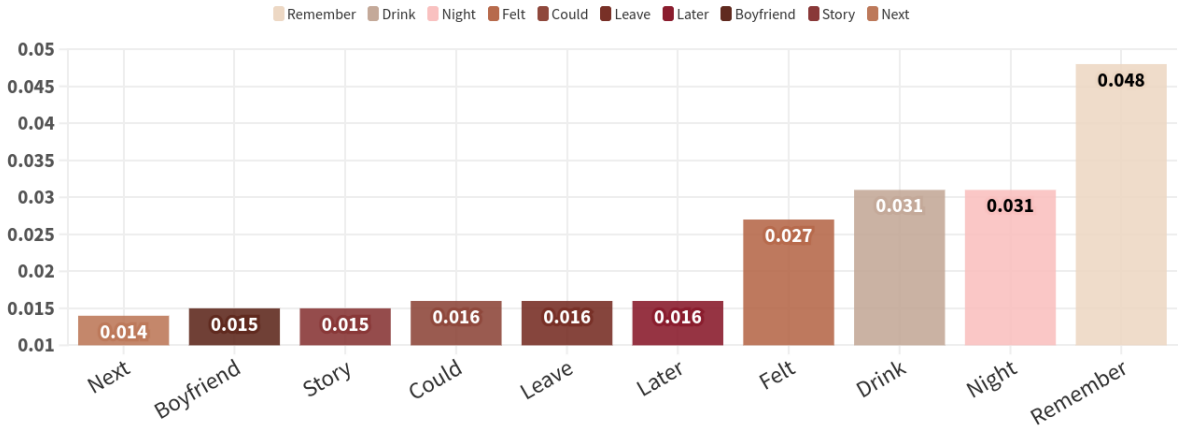


Figura 19 – Tópico 3

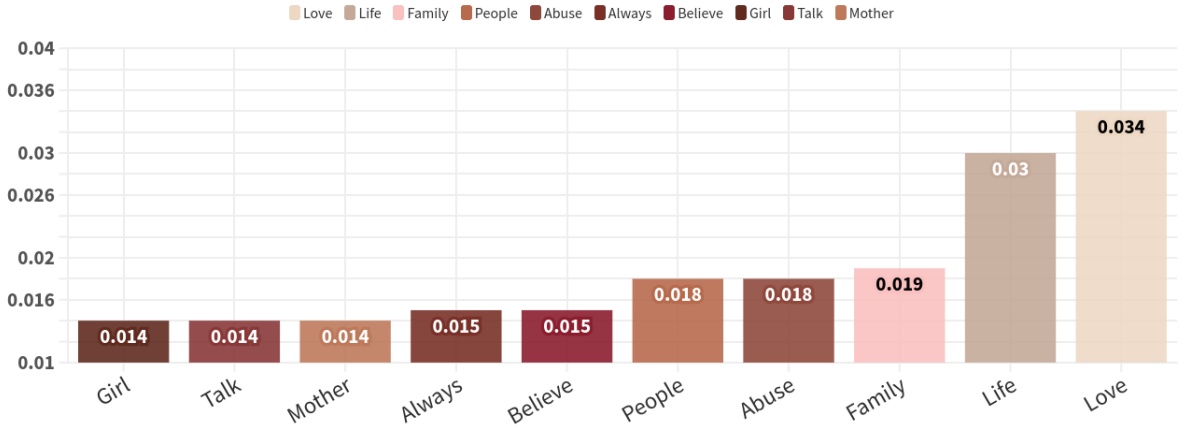


Figura 20 – Tópico 4

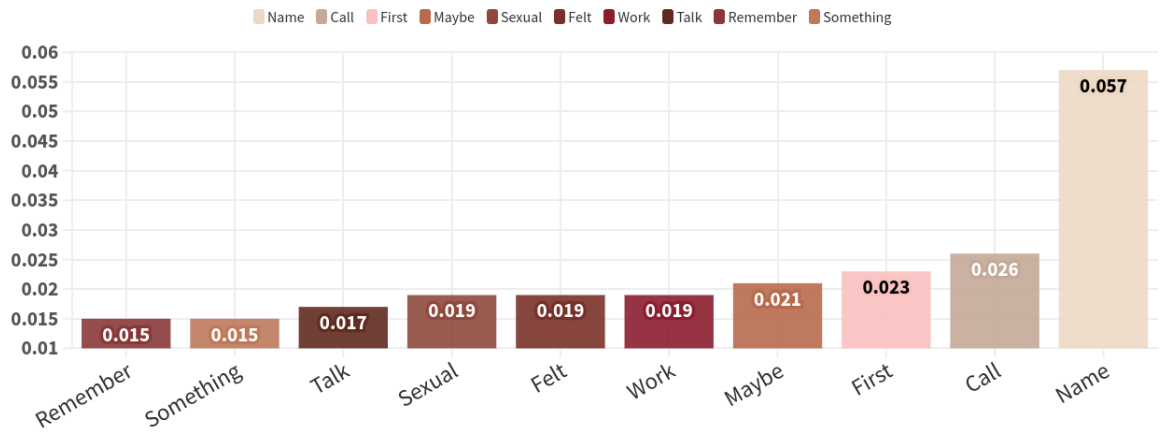


Figura 21 – Tópico 5

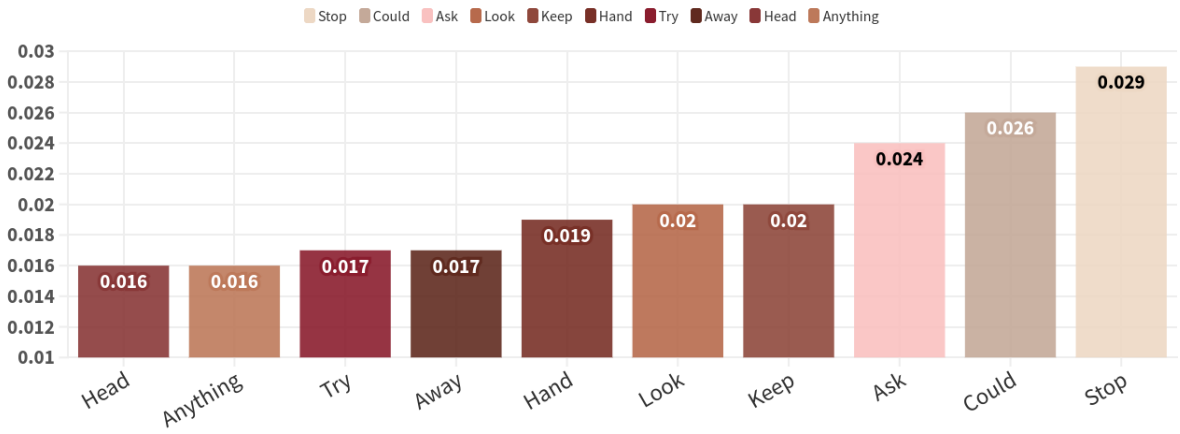


Figura 22 – Tópico 6

É perceptível que os tópicos produzidos pelo modelo nesta base de dados não conseguem ser tão bem delineados por uma área de assunto definida. Tal fato deve ocorrer pela escassez dos dados e portanto resulta numa classificação menos satisfatória em comparação à produzida pela base de dados anterior na seção 4.1.

Não obstante, ainda que não tão bem definidos, pode-se tentar descrever o assunto geral abordado por cada tópico, por exemplo, no tópico 1, representado pela Figura 17, tem-se a experiência de uma vítima de abuso em ambiente de trabalho. Já no tópico 2 na Figura 18, é abordada a temática de violência sexual cometida pelo namorado da vítima, em abiente escolar. O tópico 3 representado pela Figura 19 discorre sobre a vítima que sofre abuso sexual enquanto embriagada e não consegue se lembrar no dia seguinte do ocorrido. Na Figura 20, o tópico 4 explicita a relação de abuso em meio familiar e falta de credibilidade por parte da família. O tópico 5, na Figura 21, aborda de forma mais genérica a violência sexual da vítima que não se lembra do ocorrido. E, por fim, o tópico 6, na Figura 22, discorre sobre a vítima que tenta se desvencilhar de seu agressor, porém sem sucesso.

Com tais resultados obtidos, é possível perceber um padrão entre os assuntos presentes na base de dados de violência doméstica, o que indica a potencial capacidade de utilização desta abordagem em trabalhos futuros ou para diferentes fontes de relatos no intuito de entender e compreender melhor a violência doméstica e seus padrões, para o posterior desenvolvimento de políticas públicas de prevenção à violência.

Figura 23 – Nuvem de Palavras

Uma forma visual de analisar o peso das palavras produzidas por um modelo de Processamento de Linguagem Natural é a Nuvem de Palavras. A representação leva em consideração a importância da palavra no contexto geral, ou seja, quanto mais relevante a palavra, maior ela é representada na Nuvem. Na Figura 23 tem-se a Nuvem de Palavras dos tópicos reproduzidos pelo modelo treinado, utilizando o peso de cada palavra para representar sua relevância.

É possível observar que a maior parte das palavras mais representativas dos tópicos dizem respeito a relacionamentos pessoais como familiares ou namorados das vítimas, bem como os ambientes em que foi situada a violência, tais como ambientes de trabalho ou em festas. Foi possível identificar um padrão de palavras como "pare", "pedi"e "sair", indicando um comportamento resistivo da vítima durante o ato de violência.

4.2.1 Validação do Modelo

Em seguida, realizou-se a validação do modelo, a partir do Conjunto de Teste selecionado aleatoriamente no início do treinamento. O modelo treinado foi aplicado para os 50 documentos e foi analisada manualmente a performance deste em dados ainda não vistos. Tal abordagem é comum no contexto de Aprendizado de Máquina, pois indica o real comportamento e desempenho do modelo gerado em dados novos.

Primeiramente, autora deste trabalho classificou manualmente cada texto com os tópicos disponíveis gerados pelo modelo, como apresentados anteriormente pelas Figuras 17, 18, 19, 20, 21 e 22. Em seguida, a classificação dada pelo modelo à cada relato foi avaliada

de acordo com as 4 categorias listadas a seguir para a posterior análise de performance do modelo de acordo com os pesos designados a cada categoria:

- **Categoria 1:** o modelo classificou o texto com tópicos completamente diferentes dos tópicos rotulados manualmente - *Peso 0*
- **Categoria 2:** o modelo classificou o texto com pelo menos dois tópicos iguais aos pré-rotulados, porém com pesos errôneos - *Peso 1*
- **Categoria 3:** o modelo classificou o texto com os mesmos tópicos pré-rotulados, porém em ordens de importância diferentes - *Peso 2*
- **Categoria 4:** o modelo classificou o texto da mesma forma que a rotulação manual considerando os tópicos disponíveis para tal - *Peso 3*

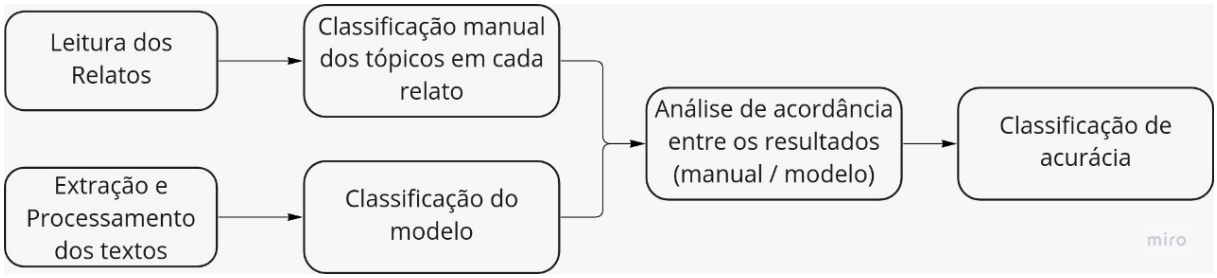


Figura 24 – Processo de Validação Manual do Modelo

Na Figura 24 foi representado o processo de análise a cálculo da acurácia do modelo, de forma a validar os resultados produzidos.

Após a classificação do Conjunto de Teste, foi calculada a média total, obtendo-se 71,8% de acurácia na performance do modelo. Fez-se também a análise do tópico mais presente na base de dados estudada e foi constatado que o tópico n^o 4 é o mais recorrente, o que indica uma maior ocorrência de abusos sexuais cometidos no contexto intrafamiliar, em comparação com outros tipos de abuso, especificamente para a base de dados utilizada nesta análise.

Considerando as limitações da base de dados e a classificação empírica realizada manualmente, a acurácia obtida foi satisfatória e dentro do esperado, o que reafirma o potencial da metodologia adotada na classificação de relatos de violência doméstica.

5 Conclusão

Historicamente, a desigualdade de gênero e a violência doméstica estiveram presentes na vida das mulheres de diversos países ao redor do mundo. De acordo com (VIOLÊNCIA..., s.d.), as relações familiares giram em torno da figura paternal masculina que representa a autoridade máxima da família; assim, a violência contra a mulher é resultado de relações de poder construídas ao longo da história pela desigualdade de gênero e consolidadas por uma ideologia patriarcal machista.

Em 1950 a Comissão sobre a Situação das Mulheres, *Commission on the Status of Women*, criada pelo Conselho Econômico e Social da Organização das Nações Unidas - *ECOSOC*, formulou tratados que afirmam que os direitos humanos deveriam ser aplicados para ambos, homens e mulheres, igualmente e sem nenhuma distinção. Porém no Brasil, apenas em 2004 foi criada a Lei Maria da Penha, que assegura à mulher o direito fundamental do ser humano de viver sem violência (ALONSO et al., 2014).

Apesar das medidas protetivas que vêm sendo implementadas desde o século passado, a violência contra a mulher continua bastante presente na sociedade atualmente. De acordo com a pesquisa *UN Women* (UN..., s.d.), no mundo, quase uma em cada três mulheres sofreu violência física ou sexual cometida por seu parceiro pelo menos uma vez na vida, resultando em 30% da população feminina com mais de 15 anos.

Recentemente, durante a pandemia do COVID-19, as ocorrências de violência cresceram consideravelmente em virtude do confinamento da vítima com seu agressor. Em (DAHAL et al., 2020), estudos demonstram que menos de 40% das mulheres que sofrem violência procuram ajuda ou reportam o ato para as autoridades locais, sendo que a maioria recorre a amigos e família porém poucas procuram instituições formais. Dahal indica que, em oito países asiáticos houve poucos boletins de ocorrência sobre violência no período da pandemia, porém pesquisas virtuais com os termos "sinais de abuso físico", "relacionamento abusivo", "definição de violência doméstica"e "como cobrir hematomas no rosto"aumentaram drasticamente: 47% na Malásia, 63% nas Filipinas e 55% em Nepal, no período entre Outubro de 2019 e setembro de 2020.

Os dados mencionados confirmam que a luta contra a violência doméstica ainda está sendo diariamente travada e que medidas protetivas às mulheres são necessárias para garantir a saúde e bem estar daquelas mais vulneráveis a esse tipo de situação. Neste contexto, este trabalho realizou uma pesquisa utilizando métodos de Extração de Dados e Aprendizado de

Máquina para recuperar relatos de vítimas e analisar os assuntos mais presentes em relatos de violência.

Como apresentado no capítulo 4, a metodologia adotada se mostrou promissora e alcançou o objetivo almejado no início deste trabalho. Apesar dos dados utilizados não serem bem estruturados e a base de dados não ser vasta o suficiente para atingir resultados tão promissores quando comparados com os resultados obtidos no teste realizado na seção 4.1, foi possível delinear os assuntos principais abordados pelos relatos com uma porcentagem de 71,8% de acurácia.

Além de comprovar a usabilidade da metodologia construída, foi possível apresentar uma noção geral dos assuntos mais presente nos relatos utilizados neste trabalho, e concluir que o tópico n^o 4 - violência familiar, representado na Figura 20, se mostra como o mais recorrente dentre os tópicos produzidos pelo modelo no contexto da base de dados utilizada, evidência que indica uma maior ocorrência de abusos sexuais e físicos no âmbito familiar da vítima em comparação com outros tipos de violência.

Tais resultados já demonstram o início da análise que pode ser feita e aplicada nas medidas de prevenção e políticas públicas de segurança, pois indicam uma possível área de enfoque que deve ser considerada na aplicação dessas medidas.

A proposta desenvolvida neste trabalho mostra-se, portanto, não apenas viável, como uma solução barata de análise de relatos de violência doméstica, que tem o potencial de contribuir e aprimorar os resultados das medidas de prevenção e segurança, dado que evidencia os assuntos e contextos mais recorrentes no âmbito de violência contra a mulher.

Trabalhos futuros podem considerar uma base diferente de relatos para posterior conclusão sobre a especificidade de cada base e como isso pode influenciar nos resultados finais da análise de tópicos, bem como a quantidade total de relatos, ou seja, a dimensão da base, pode influenciar na interpretabilidade de cada tópico.

Referências

ALGHAMDI, R.; ALFALQI, K. **A Survey of Topic Modeling in Text Mining**. v. 6. 2015. Disponível em: <www.ijacsa.thesai.org>. Citado na p. 14.

ALONSO, J.; MORERA, C.; ESPÍNDOLA, D.; CARVALHO, J. B. D.; MOREIRA, A. R.; PADILHA, M. I. **VIOLÊNCIA DE GÊNERO: UM OLHAR HISTÓRICO**. v. 5. 2014. P. 54–66. Disponível em: <http://www.abennacional.org.br/centrodememoria/here/vol5num1artigo5.pdf>. Citado nas pp. 7, 37.

BLEI, D. M.; NG, A. Y.; EDU, J. B.; JORDAN, M. I. **Latent Dirichlet Allocation**. v. 3. 2003. P. 993–1022. Citado nas pp. 17–19.

CONVENÇÃO Interamericana para Prevenir, Punir e Erradicar a Violência Contra a Mulher "Convenção de Belém do Pará"(1994). Citado na p. 7.

DAHAL, M.; KHANAL, P.; MAHARJAN, S.; PANTHI, B.; NEPAL, S. **Mitigating violence against women and young girls during COVID-19 induced lockdown in Nepal: A wake-up call**. v. 16. BioMed Central Ltd, set. 2020. DOI: 10.1186/s12992-020-00616-w. Citado na p. 37.

DRUGS, U. N. O. on; CRIMES. **Killings of women and girls by their intimate partner or other family members**. 2021. Citado na p. 8.

EXPLORING the Space of Topic Coherence Measures. In. ISBN 9781450333177. Citado na p. 19.

GALLAGHER, S.; RAFFERTY, A.; WU, A. **Stanford NLP Overview**. Disponível em: <https://cs.stanford.edu/people/eroberts/courses/soco/projects/2004-05/nlp/overview.html>. Citado nas pp. 12, 13.

GENSIM. Disponível em: <https://radimrehurek.com/gensim/>. Citado nas pp. 26, 29.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electronic Markets**, v. 31, p. 685–695, 3 set. 2021. ISSN 1019-6781. DOI: 10.1007/s12525-021-00475-2. Citado nas pp. 13, 14.

JOHRI, P.; KHATRI, S. K.; AL-TAANI, A. T.; SABHARWAL, M.; SUVANOV, S.; KUMAR, A. Natural Language Processing: History, Evolution, Application, and Future Work. In: v. 167, p. 365–375. ISBN 9789811597114. DOI: 10.1007/978-981-15-9712-1_31. Citado na p. 13.

JONES, K. S. **Natural Language Processing: A Historical Review***. Citado na p. 13.

KORENCIC, D.; RISTOV, S.; REPAR, J.; SNAJDER, J. A Topic Coverage Approach to Evaluation of Topic Models. **IEEE Access**, Institute of Electrical e Electronics Engineers Inc., v. 9, p. 123280–123312, 2021. ISSN 21693536. DOI: 10.1109/ACCESS.2021.3109425. Citado na p. 19.

LEI 11.340/2006. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11340.htm>. Citado na p. 7.

LIN, J. **On The Dirichlet Distribution**. 2016. Citado nas pp. 15, 16.

MANZOOR, M. A.; HASSAN, S.-U.; MUAZZAM, A.; TUAROB, S.; NAWAZ, R. Social mining for sustainable cities: thematic study of gender-based violence coverage in news articles and domestic violence in relation to COVID-19. **Journal of Ambient Intelligence and Humanized Computing**, abr. 2022. ISSN 1868-5137. DOI: 10.1007/s12652-021-03401-8. Citado nas pp. 20, 21.

MINKA, T. P. **Estimating a Dirichlet Distribution**. 2000. Citado na p. 15.

NATURAL Language Toolkit. Disponível em: <https://www.nltk.org/>. Citado na p. 24.

NEWS of the Brazilian Newspaper. Disponível em: <https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol>. Citado na p. 28.

OUR Wave. Disponível em: <https://www.ourwave.org/>. Citado nas pp. 10, 20, 22.

PIERCE, J. R.; CARROLL, J. B.; HAMP, E. P.; HAYS, D. G.; HOCKETT, C. F.; OETTINGER, A. G.; PERLIS, A. **LANGUAGE AND MACHINES COMPUTERS IN TRANSLATION AND LINGUISTICS A Report by the**. Citado na p. 13.

QADER, W. A.; AMEEN, M. M.; AHMED, B. I. An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. In: p. 200–204. ISBN 9781728143774. DOI: 10.1109/IEC47844.2019.8950616. Citado na p. 25.

RICHARDSON, L. **Beautiful Soup**. Disponível em: <https://beautiful-soup-4.readthedocs.io/en/latest/>. Citado na p. 23.

SEGURANÇA PÚBLICA, F. B. de. **Violência doméstica durante a pandemia de Covid-19-ed. 2**. Disponível em: <https://gazetaweb.globo.com/portal/noticia/2020/05/denuncias-de-violencia-domestica-voltam-a-subir-e-crescem-73-na-italia_105546.php>. Citado nas pp. 7, 8.

TURING, A. M. **Computing Machinery and Intelligence**. Citado na p. 12.

UN Women. Disponível em: <https://www.unwomen.org/en/what-we-do/ending-violence-against-women/facts-and-figures>. Citado nas pp. 8, 37.

VAYANSKY, I.; KUMAR, S. A. A Review of Topic Modeling Methods. **Information Systems**, Elsevier Ltd, v. 94, dez. 2020. ISSN 03064379. DOI: 10.1016/j.is.2020.101582. Citado nas pp. 14, 15, 18.

VIOLÊNCIA conjugal: problematizando a opressão das mulheres vitimizadas sob olhar de gênero. Citado na p. 37.

WILD, C. THE CONCEPT OF DISTRIBUTION. **STATISTICS EDUCATION RESEARCH JOURNAL**, International Association for Statistical Education, v. 5, p. 10–26, 2 nov. 2006. DOI: 10.52041/serj.v5i2.497. Citado na p. 15.

XUE, J.; CHEN, J.; GELLES, R. Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter. **Violence and Gender**, v. 6, p. 105–114, 2 jun. 2019. ISSN 2326-7836. DOI: 10.1089/vio.2017.0066. Citado nas pp. 19, 20.