UNIVERSIDADE DE SÃO PAULO
FACULDADE DE CIÊNCIAS FARMACÊUTICAS
Curso de Graduação em Farmácia-Bioquímica

VICTOR HUGO ALVES

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTO PARA A CLASSIFICAÇÃO E ANÁLISE TEXTUAL DE ARTIGOS CIENTÍFICOS SOBRE ANTIDEPRESSIVOS.

Trabalho de Conclusão de Curso de Farmácia-Bioquímica da Faculdade de Ciências Farmacêuticas da Universidade de São Paulo.

Orientador:

Prof. Dr. Gabriel Lima Barros de Araújo

São Paulo
2021

“Pesquisar é ver o que todos já viram
e pensar o que ninguém pensou.”

- Albert Szent-Györgyi

SUMÁRIO

Lista de abreviaturas

AAP	*Associação Americana de Psiquiatria*
BDNF	*Brain-Derived Neutrophic Factor*
BoW	*Bag-of-Words*
GUI	*Graphical User Interface*
IDF	*Inverse Document Frequency*
LDA	*Latent Dirichlet Allocation*
NMDA	*N-Methyl-D-Aspartate*
OMS	*Organização Mundial da Saúde*
TF	*Term Frequency*
TF-IDF	*Term Frequency - Inverse Document Frequency*

RESUMO

Alves, V. H. **Aplicação de técnicas de mineração de texto para a classificação e análise textual de artigos científicos sobre antidepressivos.** 2021. Trabalho de Conclusão de Curso de Farmácia-Bioquímica – Faculdade de Ciências Farmacêuticas – Universidade de São Paulo, São Paulo, 2021.

Palavras-chave: Antidepressivos, cetamina, modelagem matemática, agrupamento

INTRODUÇÃO: Depressão é uma doença comum definida por um sentimento de tristeza, baixa autoestima e diminuição da percepção de valor próprio que pode levar ao suicídio. Segundo a OMS é estimado que 3,8% da população mundial sofra com depressão. Para o tratamento da depressão temos a terapia cognitiva comportamental, terapia eletroconvulsiva e terapias farmacológicas, originalmente atuando sobre a recaptação de monoaminas, como serotonina, noradrenalina e dopamina, e também terapias novas inibindo receptores glutamatérgicos e estimulando a neurogênese. Por ser um campo de estudo vasto, o número de publicações sobre o tema de depressão e antidepressivos vem crescendo consideravelmente. E devido a expansão rápida da quantidade de conteúdo, o uso de técnicas de mineração de texto pode ser bastante útil para tornar análises textuais mais dinâmicas e adicionar valor através de novas perspectivas baseadas em dados. **OBJETIVOS:** Realizar um agrupamento por técnicas de modelagem de texto e buscar novas relações entre fármacos antidepressivos **MATERIAIS E MÉTODOS:** Foi realizada uma extração de 395 artigos científicos disponíveis para consulta dentro da base de dados da Elsevier, processos de limpeza, preparação e organização dos dados antes da criação dos modelos de K-médias e LDA. Após estes processos dois métodos de agrupamentos foram realizados no *software* e linguagem de programação R versão: 4.0.3 aplicados com validação cruzada prévia para encontrar os pontos ótimos e depois comparar ambos os modelos. **RESULTADOS:** O modelo LDA apresentou ótimo desempenho no processo de agrupamento dos artigos científicos, sendo observado a partir da validação cruzada que 50 tópicos foi o ideal. Depois de agrupados, os tópicos foram avaliados e cada um deles recebeu um tema. Os artigos científicos que foram alocados dentro dos temas relacionados a cetamina foram avaliados sobre sua proximidade a temas adjacentes, procurando por intersecções nas áreas associadas a cetamina, em especial sugerindo o estudo do impacto neurológico da cetamina em mães e nas gerações seguintes. **CONCLUSÃO:** Apesar das limitações apresentadas por este trabalho em relação ao número de documentos analisados e o período restrito a 2020 e 2021, o uso da técnica de LDA trouxe uma visão e organização dos artigos científicos bastante interessante e permitiu que uma nova indicação para área de estudo sobre a cetamina.

1. Introdução

1.1. Depressão

Segundo a OMS, a depressão é uma doença comum e é estimado que 3,8% da população mundial esteja afetada (OMS, 2021). A depressão é uma doença incapacitante que difere de tristeza devido a sua duração prolongada, baixa na auto estima, diminuição na percepção do valor próprio (AAS, 2021).

A depressão pode ser avaliada a partir de diferentes hipóteses. A hipótese monoaminérgica da depressão é bastante aceita e sugere que a depressão é causada pela redução dos neurotransmissores serotonina, dopamina e noradrenalina (FREIS, 1954). Uma outra hipótese que tem bastante relevância para a depressão é a hipótese neurotrópica da depressão que descreve que o estresse causa mudanças como redução do volume de algumas regiões cerebrais como o córtex pré-frontal e o hipocampo, inibição de neurogênese e tamanho dos dendritos (DUMAN; LI, 2012).

Existem diferentes tratamentos para depressão e eles podem ser administrados sozinhos ou em paralelo. Os tratamentos podem ser divididos em dois grupos: Não farmacológicos e farmacológicos. Para os não farmacológicos os tratamentos incluem: Terapia Cognitiva Comportamental e Terapia Eletroconvulsiva. Para os farmacológicos temos os fármacos antidepressivos (AAP, 2021).

1.2. Antidepressivos

Os antidepressivos formam a classe farmacológica que atua reduzindo os sintomas da depressão e de outras doenças relacionadas, como a ansiedade. Os inibidores da recaptação de serotonina são os antidepressivos de primeira linha para o tratamento da depressão. Estes antidepressivos têm como seu mecanismo de ação o bloqueio da recaptação de serotonina (NUTT et al., 1999), bloqueio da recaptação de noradrenalina, inibidores da enzima monoamino oxidase e agonistas de receptores de serotonina (KRISHNAN; NESTLER, 2008). Aproximadamente dois terços dos pacientes apresentam uma melhora significativa em seu tratamento, porém metade destes pacientes necessitam de reavaliação clínica geralmente associada a administração de um novo antidepressivo. Os sintomas começam a diminuir significativamente nestes

pacientes somente após um período médio de 2 a 3 semanas de tratamento (KRISHNAN; NESTLER, 2008). O terço final do total apresenta uma condição especial dentro da depressão conhecida como depressão resistente ao tratamento. Ou seja, estes pacientes não respondem bem aos antidepressivos clássicos, mesmo após o uso de diferentes antidepressivos (TRIVEDI *et al.*, 2006).

Para este trabalho vamos olhar também para as publicações sobre antidepressivos e estas aumentaram consideravelmente através dos anos. No site Web of Science, podemos ver um agregado de publicações de diversos autores, categorias, revistas e editoras. Quando procuramos por “Antidepressants” e filtramos o período entre 2000 e 2020, observamos que o número de publicações anuais passou de 1.627 para 4.376. Este é um crescimento de 170% em 20 anos e um total de 64 mil publicações (WoS, 2021).

Para analisar o conteúdo presente em tantos artigos científicos o uso de ferramentas computacionais e de modelagem matemática se tornam essenciais para adicionar direcionamento e trazer novas visões aos tópicos a serem analisados.

1.3. Ferramentas computacionais de mineração de texto para a classificação e análise textual

Avanços recentes em ferramentas computacionais têm permitido o processamento de enorme e crescente quantidade e de informações disponíveis em documentos textuais, por exemplo na área da saúde (CHENG et al, 2010). Entre eles destacam-se o uso de Big Data, a aprendizagem de máquina e o uso de validação cruzada, bem como as abordagens de agrupamentos e o processamento de linguagem natural.

1.3.1. *Big Data*, aprendizagem de máquina e validação cruzada

Big Data é um termo utilizado para descrever o conjunto de informações que existe e está sendo gerada agora. Procurando melhorar e otimizar a busca e compreensão de diferentes conhecimentos, técnicas de análise estatística são empregadas a essas enormes quantidades de texto a fim de classificá-las em função de seu conteúdo.

A aprendizagem de máquina nasceu de teorias de reconhecimento de padrões e do objetivo de desenvolver computadores capazes de aprender sem precisarem ser programados para gerar esses resultados. Para alcançar tal objetivo, os dados fundamentam esse processo, pois ao expor modelos a novos dados é possível criar um processo de adaptação independente (SAS, 2021). A validação cruzada é uma classe de métodos utilizados para verificar a capacidade de generalização de um dado modelo através da reamostragem, evitando sobre-ajuste (BERRAR, 2019).

Um grande problema presente na criação de modelos matemáticos é que na busca de aumentar a acurácia do modelo, este fica muito específico ao grupo de dados de treinamento e perde a capacidade de prever novos dados. A esse evento é dado o nome de sobre-ajuste (BERRAR, 2019). Um dos métodos de validação cruzada é conhecido como k partições, onde os dados de treinamento são subdivididos em k subgrupos de tamanhos similares. O modelo então é treinado em k-1 subgrupos e o último grupo é então utilizado para testar a acurácia do modelo. Como o modelo nunca interagiu com um dos subgrupos durante a fase de treinamento, a acurácia observada é uma medida mais próxima da capacidade de generalização do modelo (BERRAR, 2019).

Este processo pode ser repetido dentro das possíveis combinações entre subgrupos de treinamento e subgrupos de teste, permitindo que uma acurácia média seja encontrada, diminuindo a possibilidade de uma medida distorcida representar o modelo (BERRAR, 2019).

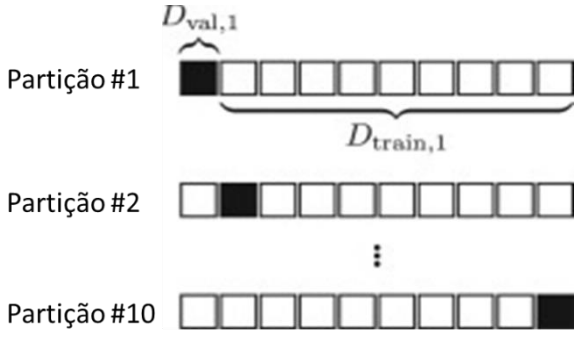


Figura 1: Exemplo visual da técnica de k partições, neste caso com k = 10. Cada quadrado preto representa o subgrupo de avaliação (ou de teste) e os quadrados brancos representam os subgrupos de treinamentos (Berrar, 2019 adaptado).

1.3.2. Agrupamentos

Segundo Jain et al. o processo da análise de dados pode ser dividido entre exploratório ou confirmatório, mas independente da natureza do processo, o agrupamento ou classificação de dados é um elemento chave para a análise de dados de forma geral. Os agrupamentos podem ser baseados em agrupamentos “naturais”, inerentes aos dados, como pessoas de um mesmo país falarem um mesmo idioma ou por quão bem um dado modelo pode classificar as informações (JAIN; MURTY; FLYNN, 1999).

O processo de agrupamentos busca encontrar padrões de semelhança ou proximidade nos dados, geralmente pontos em um espaço multidimensional, frente a sua distribuição. A definição de semelhança ou proximidade pode ser explorada de formas distintas dependendo da lógica aplicada e graças a essa diferença de interpretação várias técnicas de agrupamentos surgiram (JAIN; MURTY; FLYNN, 1999).

As técnicas de agrupamentos são todos métodos de redução dimensional e não supervisionados. Um processo paralelo existe, mas de natureza supervisionada e essas técnicas são conhecidas como análise discriminante. Ambos os métodos de análise têm o mesmo objetivo, porém seguem por caminhos distintos para gerar os agrupamentos (JAIN; MURTY; FLYNN, 1999).

Segundo Jain et al. é possível generalizar o processo de agrupamento a 5 principais passos (3 essenciais e 2 opcionais), como vemos a seguir:

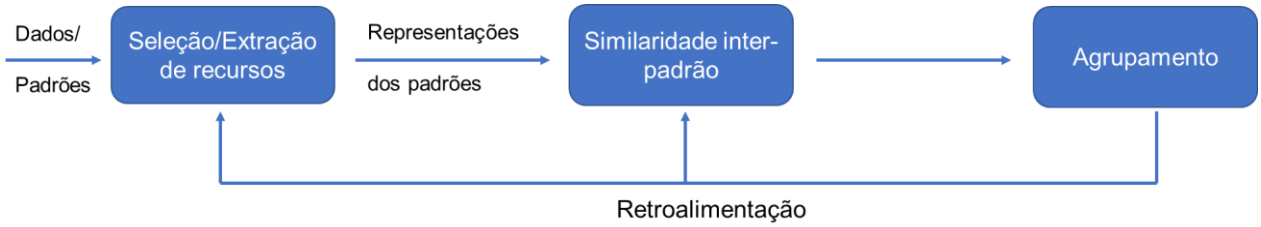


Figura 2: Sequência dos 3 passos essenciais utilizados pelas técnicas de agrupamento. Fonte: Modificado de Jain et al., 1999.

Seleção de recursos é o ato de selecionar quais variáveis são mais relevantes dentro do processo de agrupamento. Algumas variáveis são combinações lineares de

variáveis anteriores ou aumentar o erro associado a técnica. Já a Extração de recursos é definida por qualquer transformação dos dados originais em novos dados, como por exemplo normalização dos dados (JAIN; MURTY; FLYNN, 1999).

Similaridade inter-padrão é a distância observada entre os dados ou entre grupos de dados e um novo ponto. Um exemplo comum de distância utilizada é a distância Euclidiana, mas também existe a distância de Mahalanobis, frequentemente utilizada para avaliar agrupamentos de natureza hiper elipsoide (JAIN; MURTY; FLYNN, 1999).

Agrupamento pode apresentar 2 tipos de resultados. Eles podem ser agrupamentos duros, do inglês *hard clustering*, ou seja, cada observação é classificada como pertencente a um único grupo ou agrupamentos vagos, do inglês *fuzzy clustering* ou *soft clustering*, onde cada observação tem um grau variável de pertencimento a cada um dos agrupamentos (JAIN; MURTY; FLYNN, 1999).

Outro ponto importante dentro de qualquer análise de agrupamento é a definição dos números de grupos presentes dentro do conjunto de dados. Começando com agrupamentos hierárquicos, o processo parte de um único grupo contendo todos os dados e vai criando segregações até chegar em unidades únicas de dados, ficando nas mãos do analista a decisão e uso de outros conhecimentos sobre a natureza dos dados para definir onde é o melhor ponto para definir o número de grupos (ANDERBERG, 1973).

Em muitos casos o uso da validação cruzada é uma forma bastante efetiva de dar uma visualização da organização dos dados. Diferentes métodos possuem diferentes métricas para avaliar o ganho de explicação entre usar dado número de grupos. A avaliação de efeito de diferentes números de grupos na variância explicada ou erro associado a técnica geralmente são utilizados para selecionar o número de grupos, como veremos neste trabalho. Além disso, igualmente importante é o fato de que o número de grupos selecionados deve fazer sentido quando avaliado, ou seja, quando agrupados os dados trazem uma nova informação que anteriormente estava “escondida” pela alta entropia presente no conjunto de dados (FAWCETT, 2013, p. 49-56; ANDERBERG, 1973, p.21-23).

1.3.3. Processamento de Linguagem Natural – *Natural Language Processing* (NLP)

O processamento de linguagem natural, ou NLP como é mais conhecido, é um conjunto de métodos para tornar as linguagens humanas acessíveis a computadores. Esta área de pesquisa surgiu da união de diferentes áreas de conhecimento, em especial linguística e estatística (EISENSTEIN, 2018, p.1-3).

Podemos notar a presença destes métodos por toda a internet, como algoritmo de tradução, classificação de e-mails e também em cada uma das sugestões e buscas realizadas pelo site da Google, por exemplo (EISENSTEIN, 2018, p.1-3).

Para desenvolver estes métodos, a pesquisa de NLP se baseia em técnicas de aprendizagem de máquina, mas pela natureza dos dados utilizados, algumas particularidades ocorrem durante análises de NLP. Por exemplo, as palavras ou combinações de palavras a serem analisadas são observações discretas. Outro ponto relevante é que a distribuição das palavras em documentos não segue uma distribuição normal, mas uma distribuição Zipf ou distribuição *Power law* ou distribuição de potência, ou seja, algumas palavras aparecem muito e a frequência das palavras diminui exponencialmente, logo é possível ver uma tendencia linear apenas após uma conversão da escala para seu logaritmo (EISENSTEIN, 2018, p.1-3).

Um exemplo de aprendizagem de máquina aplicados à área da saúde foi publicado por Lionel Cheng e Bradley Erickon. Utilizando uma técnica chamada *Support Vector Machine* (SVM), os pesquisadores classificaram e indicaram a progressão de tumores cerebrais a partir de laudos de exames de ressonância magnética. Esta técnica pertence ao grupo de aprendizado de máquina supervisionado, o qual necessita necessariamente que os dados de treinamento e de avaliação apresentem uma variável resposta categórica ou contínua. Para atender essa necessidade, os dois pesquisadores principais classificaram os textos usados para treinamento e avaliação do modelo (CHENG et al, 2010).

1.3.4. Modelagem de tópicos

A modelagem de tópicos se encontra dentro da divisão de aprendizado de máquina não supervisionada, ou seja, não é necessário que as observações apresentem uma variável resposta/variável dependente durante a fase de treinamento e nem de avaliação. Este também é um método probabilístico, baseando-se na distribuição de probabilidade conjunta de suas variáveis, ou seja, avaliando a probabilidade de 2 ou mais eventos observados, no caso específico deste trabalho os eventos são as palavras, ocorrerem ao mesmo tempo (ASMUSSEN; MØLLER, 2019).

A modelagem de tópicos calcula a probabilidade de textos diferentes pertencerem a um mesmo grupo através das frequências das palavras encontradas dentro dos textos. Dessa forma, textos que apresentam as mesmas palavras em grandes quantidades em seu corpo, estão relacionados entre si. Enquanto textos com palavras distintas são agrupados em diferentes tópicos (ASMUSSEN; MØLLER, 2019).

2. Objetivos:

Este trabalho tem como seu principal objetivo aplicar técnicas de análise textual em artigos científicos sobre antidepressivos, de modo a trazer novas perspectivas sobre as informações atualmente presentes na literatura dessa classe farmacológica e identificar possíveis relações entre as informações dispostas na literatura.

2.1. Objetivos específicos

Descobrir relações entre artigos científicos sobre antidepressivos através de técnicas de agrupamentos k-médias e alocação de dirichlet latente.

3. Materiais e Métodos:

3.1. Ferramentas Computacionais

Todas as análises estatísticas foram feitas utilizando a linguagem de programação R (encontrado em: < https://www.r-project.org/>), sendo parte através do GUI RStudio (encontrado em: < https://www.rstudio.com/>) e parte no Google Colaboratory (encontrado em: < https://colab.research.google.com/>. Além disso, para processos de tabulação de dados o Microsoft Excel 2016.

3.2. Extração

Para a obtenção da primeira base de dados, um script em R foi escrito para buscar o termo "*Antidepressants*" entre janeiro de 2020 e fevereiro de 2021 no site Web of Science. Dentro desta seleção, foram encontrados 64 mil registros de artigos científicos disponíveis. O script selecionou uma opção de exportar dados dos artigos disponíveis na busca, processando os dados de 500 em 500 conforme a limitação estabelecida pelo próprio site.

Após 128 ciclos de exportações, os arquivos foram unidos em uma grande base de dados inicial contendo as informações disponíveis dos 64 mil artigos sendo: O título da obra, os autores, ano de publicação, editora, área de pesquisa, DOI e entre outros.

3.3. Limpeza da extração

A limpeza dos dados se iniciou pela coluna contendo os DOIs dos artigos de interesse. Apesar da primeira base de dados formada apresentar 64 mil linhas, quando foram filtrados os DOIs únicos, apenas 16.238 linhas ficaram remanescentes, ou seja, pouco mais de um quarto da quantidade de linhas original.

3.4. Coleta dos dados

Agora com mais de 16 mil DOIs disponíveis, a estratégia original era utilizar um API, como o do Web of Science ou da Elsevier, no entanto durante o progresso do trabalho ambos os APIs não estavam retornando os documentos dos artigos.

Sendo assim, uma seleção dos DOIs disponíveis da editora Elsevier, devido a sua confiabilidade, foram selecionados e passados a um arquivo de Excel. Com os DOIs e a

planilha do Excel, foram tabelados 395 artigos científicos separados em: Introdução, Materiais e Métodos, Resultados, Discussão e Conclusão. Essa divisão se provou útil não apenas por classificar as diferentes partes dos artigos, mas também porque cada célula do Excel limita o número de caracteres a 32.767 e diversos artigos superaram essa quantidade de caracteres.

3.5. Métodos estatísticos

Todos os cálculos realizados neste trabalho foram no software e linguagem de programação R através do site do Google Colab.

Para garantir a reprodutibilidade dos resultados, a “semente” 123 foi utilizada antes de cada um dos cálculos. (set.seed(123)).

3.5.1. Método K-médias

O método de k-médias é um processo iterativo que atribui o número definido de grupos a um mesmo número de pontos de forma aleatória e um centro, ou ponto médio, é calculado para cada grupo. Cada ponto é alocado ao grupo mais próximo e depois o ponto médio é recalculado em função das posições de cada cluster (EISENSTEIN, 2018, p.96-97; ANDERBERG, 1973, p.162-163).

Devido à natureza aleatória da primeira formação de grupos, o método de k-médias não necessariamente produzirá um resultado com o mínimo de variância explicada possível em uma primeira avaliação (EISENSTEIN, 2018, p. 97). Outro ponto relevante sobre o método de k-médias é focado em minimizar a soma dos quadrados dos erros, aumentando por consequência a soma dos quadrados dos grupos.

3.5.2. Método Bolsa de palavras

Para avaliar matematicamente corpos textuais é importante extrair alguma métrica. Para tanto, precisamos realizar algumas adaptações para extrair numericamente o sentido presente dentro dos textos da forma mais eficiente possível.

A técnica de bolsa de palavras é uma forma inicial desenvolvida para tratar deste desafio. Ao utilizar essa técnica, qualquer documento é reduzido a um conjunto de palavras ou também chamado de termo ou token, numa matriz com documentos num

eixo e os tokens em seu outro eixo, ignorando o impacto da ordem das palavras, estrutura da sequência, regras gramaticais e pontuações (FAWCETT, 2013, p.252).

A matriz formada pela bolsa de palavras nos trouxe uma informação de presença ou ausência dos tokens dentro dos documentos. Essa informação nos permite gerar uma primeira classificação entre os documentos, mas ainda não nos permite avaliar a importância dos tokens entre si (FAWCETT, 2013, p. 252).

1) “Eu gosto da cor azul.” -> “eu”, “gosto”, “da”, “cor”, “azul”
2) “O céu é azul.” -> “o”, “céu”, “é”, “azul”

	azul	céu	cor	da	é	eu	gosto	o
Documento 1	1	0	1	1	0	1	1	0
Documento 2	1	1	0	0	1	0	0	1

Figura 3: Exemplo de matriz de documentos e termos. Fonte: Figura do autor.

Na figura acima, podemos ver um exemplo simples onde cada documento é fragmentado em suas palavras individuais e as palavras únicas são dispostas numa matriz para indicar a sua presença ou ausência nos documentos que compõem o corpo textual.

3.5.3. Método Frequência dos termos – Term Frequency (TF)

A frequência das palavras presentes dentro de cada documento é uma forma inicial de extrair informações. A princípio se fosse apenas indicada a presença ou ausência de palavras dentro dos documentos presentes na análise, teríamos uma escala nominal (ANDERBERG,1973, p. 27-28).

Através de uma escala nominal podemos iniciar o processo de classificação entre observações, no case deste trabalho documentos, mas este é o mínimo de informação que pode ser extraído da técnica de Bolsa de palavras (FAWCETT, 2013, p. 252).

O segundo passo seria contar a frequência de cada palavra dentro de cada documento e adicionar essa informação na matriz de documentos e tokens. É possível utilizar a frequência absoluta ou a frequência relativa ao número de palavras dentro de um mesmo documento para compor a matriz de frequência dos termos. O segundo caso é especialmente relevante quando os documentos tem tamanhos muito diferentes e um

token poderia ter maior impacto simplesmente pela magnitude da frequência absoluta dele (FAWCETT, 2013, p. 254).

1) Eu gosto da cor azul. -> “eu”, “gosto”, “da”, “cor”, “azul”
2) A casa é azul no mesmo tom de azul do céu. -> “a”, “casa”, “é”, “azul”, “no”, “mesmo”, “tom”, “de”, “azul”, “do”, “céu”

	a	azul	casa	céu	cor	da	de	do	é	eu	gosto	mesmo	no	tom
Documento 1	0	1	0	0	1	1	0	0	0	1	1	0	0	1
Documento 2	1	2	1	1	0	0	1	1	1	0	0	1	1	1

Figura 4: Exemplo de matriz de frequência dos termos nos diferentes documentos. Fonte: Figura do autor.

3.5.4. Método Frequência inversa nos documentos – Inverse Document Frequency (IDF)

A avaliação das frequências permitiu que uma nova profundidade fosse adicionada a informação dos documentos e seus tokens. No entanto, quando as frequências dos tokens estão em seus extremos, ou seja, quando algum token aparece poucas vezes ou muitas vezes em todos os documentos, seu impacto no processo de classificação é negativo (FAWCETT, 2013, p. 254-255).

Quando pouco presente, provavelmente o token não será tão relevante para o agrupamento dos documentos e por outro lado quando o token aparece muito em todos os documentos ele se torna irrelevante já que todos os documentos o têm e nenhuma distinção pode ser feita entre eles. Para ganhar eficiência nas análises, geralmente limites inferiores e superiores são aplicados ficando a critério do analista decidir quais serão estes limites. (FAWCETT, 2013, p. 254-255).

Além de limites inferiores e superiores, outra forma empregada para analisar essa relação de um termo com a sua presença nos diferentes documentos se da pela distribuição do termo nos documentos.

Esta proposta foi desenvolvida por Karen Sparck Jones em 1972, quando ela trabalhou desenvolvendo pesos para serem aplicados a palavras numa análise de sua presença ou ausência dentro de cada um dos documentos e depois comparando essa presença em relação à quantidade total de documentos analisados (Sparck-Jones, 1972). Algum tempo após esta publicação, o termo “especificidade” foi substituído por Frequência Inversa nos Documentos (ROBERTSON, 2004).

Essa presença é calculada através da seguinte equação:

$$IDF(termo) = 1 + \log\left(\frac{Número\ total\ de\ documentos}{Número\ de\ documentos\ contendo\ o\ termo}\right)$$

Através da função apresentada acima e da imagem esta é uma forma de visualizar se uma palavra é rara, ou seja, aparece em poucos documentos ou se ela está bem distribuída por todos os documentos que compõem o corpo textual (FAWCETT, 2013, p. 254-255).

O racional por trás desta técnica é de que um termo que ocorre em muitos documentos não é capaz de trazer um grande impacto em um processo de classificação, como comentado anteriormente, logo deve receber um “peso” menor (SPARCK-JONES, 1972; ROBERTSON, 2004).

3.5.5. Método Frequência dos termos-Frequência Inversa nos Documentos – Term Frequency-Inverse Dovument Frequency (TFIDF)

A interpretação de TF-IDF é o produto entre a frequência dos termos e a frequência inversa de um termo nos documentos analisados. Neste caso, a frequência dos termos descreve uma estimativa da probabilidade de ocorrência de um certo termo dentro dos textos (AIZAWA, 2003).

Já a porção IDF, ou frequência inversa nos documentos pode ser interpretada como a “quantidade de informação” descrita pelo logaritmo do inverso da probabilidade, ou frequência de presença no todo (AIZAWA, 2003).

$$TFIDF(t,d) = TF(t,d) * IDF(t)$$

Dessa forma, a técnica TF-IDF é definida como uma equação que relaciona a probabilidade de um dado termo e um peso referente a quantidade de informação que ele representa dentro do todo (AIZAWA, 2003).

3.5.6. Alocação de dirichlet latente – Latent Dirichlet Allocation (LDA)

A alocação de dirichlet latente é uma técnica de redução dimensional utilizada para generalizar textos a partir de um modelo de predição probabilístico. Dentro deste método, um grande corpo textual é formado a partir de todos os documentos a serem analisados e dentro desses documentos, são encontradas unidades ainda menores para serem feitas as estimativas e classificações. Essas unidades geralmente são palavras, mas também podem ser combinações de 2 ou mais palavras (BLEI; NG; JORDAN, 2003).

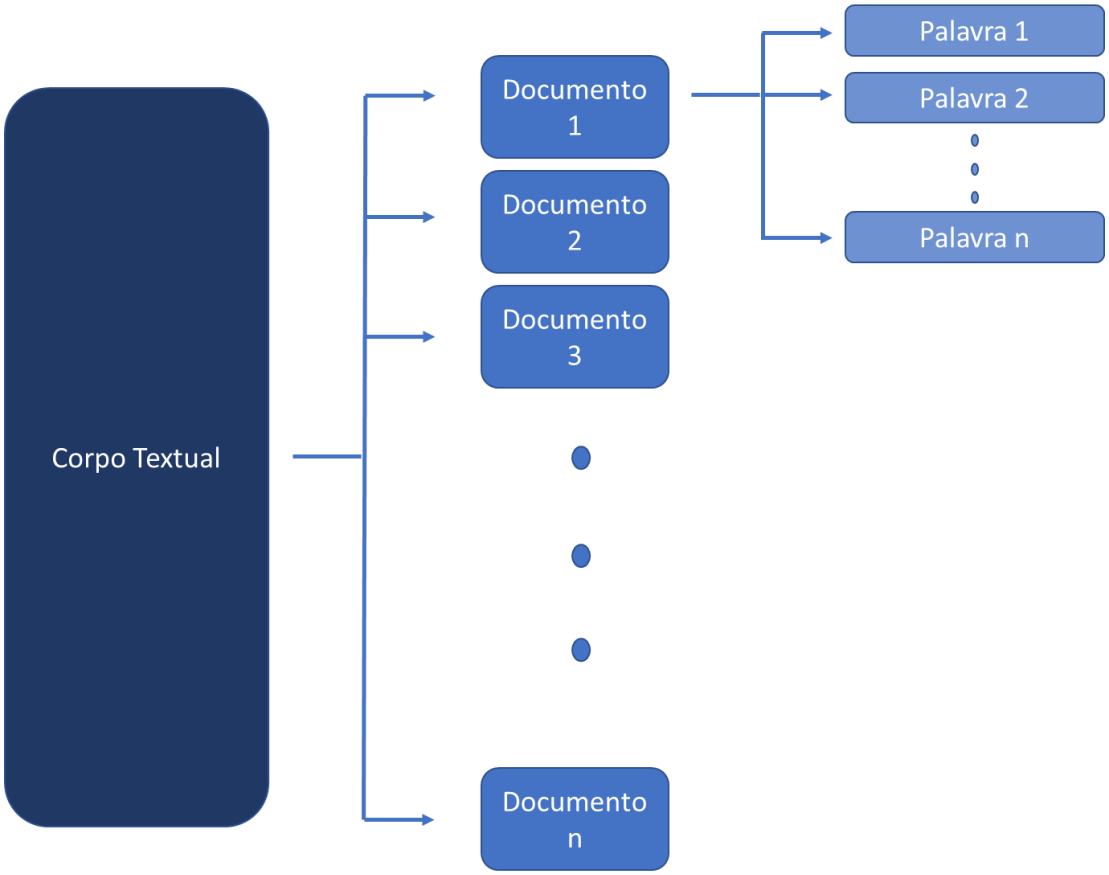


Figura 5: Gráfico das segmentações realizadas para a LDA. Fonte: Figura do autor.

Para iniciar a generalização do conteúdo textual utilizando a LDA são feitas algumas simplificações, como: A ordem das palavras dentro do corpo textual não afeta o resultado final e é assumido que os diferentes documentos textuais individuais em si também são independentes entre si. Sabemos que isso não é verdade, especialmente

na primeira especificação sobre as ordens das palavras, mas é uma necessidade para este tipo de generalização (BLEI; NG; JORDAN, 2003).

Outro ponto importante para ser ressaltado é que a simplificação acima não deve ser interpretada como independência de todas as palavras e/ou corpos textuais, bem como que eles estejam igualmente distribuídos (BLEI; NG; JORDAN, 2003).

Para averiguar as probabilidades e calcular as estatísticas dos termos dentro dos documentos, o modelo conhecido como Bag of words, é utilizado, onde cada palavra é assumida como um termo independente e será utilizado para inferir as probabilidades deste termo através das frequências observadas deste termo em relação aos outros termos em todos os documentos analisados (BLEI; NG; JORDAN, 2003).

Dessa forma, o LDA consiste de um modelo probabilístico do tipo modelo mistura, ou seja, as distribuições individuais se apresentam como sub populações da distribuição do material completo (BLEI; NG; JORDAN, 2003).

O LDA se apresenta mais versátil do que outras técnicas de redução dimensional e classificação, pois as suas estimativas não impedem que um documento seja alocado apenas a um tópico, mas apresenta diferentes tópicos em graus diferentes (BLEI; NG; JORDAN, 2003).

Esta característica de “soft classification” é o que torna o LDA mais versátil do que uma classificação Bayesiana geral, já que nesta todas as palavras de um documento são consideradas de apenas uma classe (GRIFFITHS; STEYVERS, 2004).

3.5.6.1. Perplexidade

A perplexidade é uma métrica utilizada para a avaliação de modelos de previsão de sequências, sendo um exemplo modelagem de tópicos. A perplexidade é definida como a média geométrica do inverso das probabilidades estimadas das palavras encontradas dentro daquele corpo textual (KOBAYASHI, 2014).

$$PP := \left(\prod_{w \in \mathbf{w}_\tau} \frac{1}{p(w)} \right)^{\frac{1}{N_\tau}}$$

Figura 6: Expressão matemática utilizada para calcular a perplexidade de um dado modelo. Fonte: Kobayashi, 2014.

Intuitivamente, a perplexidade de um modelo indica quantas possibilidades existem para estimar a próxima palavra num corpo de texto de teste. Logo, quanto menor a perplexidade, melhor a capacidade de generalização de um dado modelo (KOBAYASHI, 2014).

3.5.6.2. Beta

O hiper parâmetro beta descreve as probabilidades de uma dada palavra, presente dentro de cada documento pertencer a cada tópico (BLEI; NG; JORDAN, 2003).

3.5.6.3. Alfa

Alfa é o hiper parâmetro que varia entre 0 e 1. Este é passado a função gamma, que é uma função fatorial, como um de seus parâmetros para calcular a probabilidade de os tópicos estarem presentes em cada um dos documentos (BLEI; NG; JORDAN, 2003).

4. Resultados e Discussão

4.1. Avaliação geral

4.1.1. Matriz de termos

Após a remoção das palavras genéricas, pontuações, números e unidades dimensionais que apesar de terem valor no idioma falado apresentam ganho de sentido a análise, foi criada uma matriz entre termos e sua frequência no corpo textual inteiro.

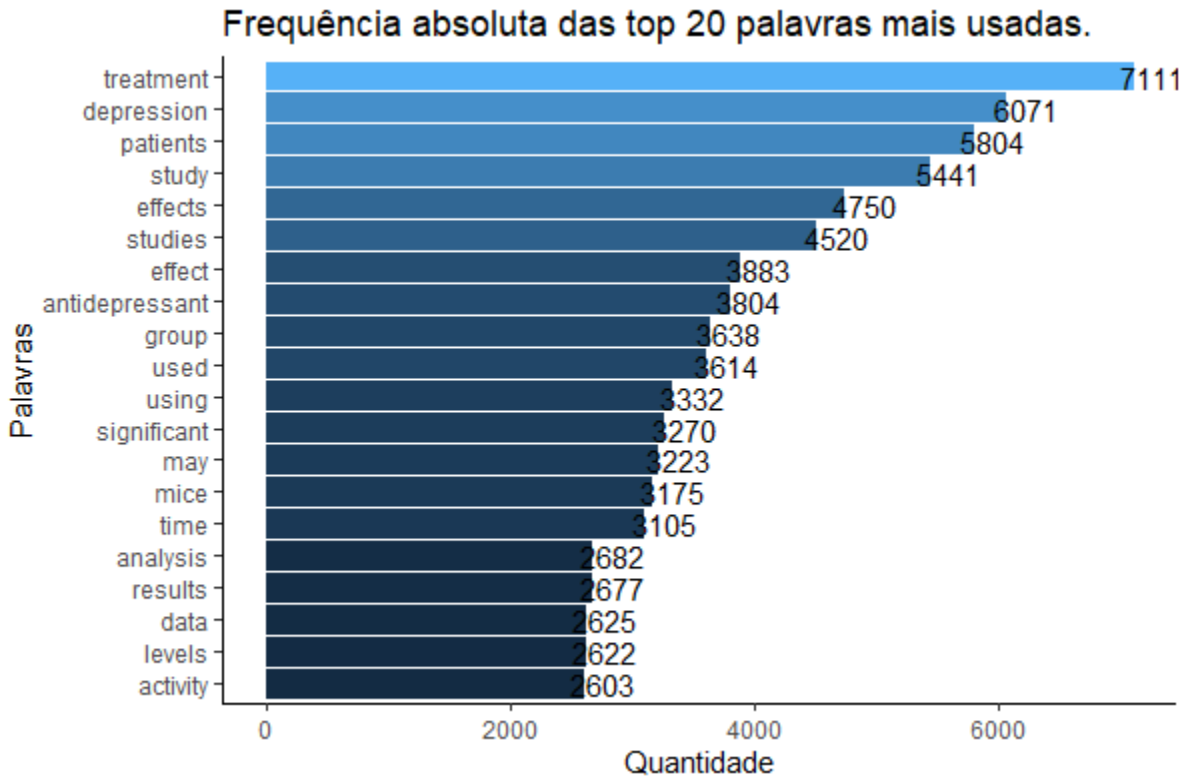


Figura 7: Top 20 palavras mais frequentes nos 395 artigos científicos. Fonte: Figura do autor.

Figura 8: Visualização das top 50 palavras no formato de “nuvem de palavras” onde a quantidade de vezes que a palavra aparece define o seu tamanho. Fonte: Figura do autor.

As duas figuras acima demonstram respectivamente as 20 e 50 palavras mais frequentes após a remoção das *stopwords*. Neste primeiro momento existem 48.732 mil palavras únicas, mas muitas dessas palavras são usadas apenas uma vez e não trazem valor ao modelo.

As palavras mais frequentes indicam que o foco dos trabalhos são tratamentos para a depressão, como é de se esperar, porém os dados ainda estão pouco organizados para que tenhamos alguma conclusão mais profunda sobre o material coletado.

4.2. Agrupamentos

Uma forma interessante de entender a natureza dos dados e como eles se relacionam é utilizando técnicas de agrupamento. Como os dados não apresentam uma classificação prévia, abordagens não supervisionadas são as mais indicadas.

Inicialmente a técnica de k-médias foi escolhida para realizar uma abordagem mais veloz e permitir que fosse retirado algum insight sobre os “grupos naturais” da informação presente.

4.2.1. K-Médias

A técnica de k-médias recebeu uma matriz contendo os documentos e as frequências das palavras dentro de cada documento. Após a limpeza dos dados, sobraram cerca de 5 mil palavras presentes na matriz original, no entanto 97% dessa matriz ainda se encontrava vazia, ou seja, sem presença dessas palavras na maioria dos artigos científicos. Logo, um ajuste foi feito removendo as palavras com maiores quantidades de 0 presentes na matriz até que chegamos a uma nova matriz com 269 palavras e apenas 31% das células da matriz estavam vazias.

Como discutido anteriormente, para o método de k-médias, uma das informações que necessitam ser fornecidas é o número de grupos que será utilizado. Para descobrir um número de grupos satisfatório para os dados da matriz acima a técnica foi realizada variando o número de grupos entre 1 e 100, aumentando um grupo de cada vez. Para cada grupo foi calculada a soma de quadrados dos grupos e a soma de quadrados totais, chegando assim à variância explicada apresentada por cada grupo. A variância explicada e o número de grupos foram dispostos num gráfico para melhor avaliação do ganho de variância explicada de acordo com cada grupo adicionado.

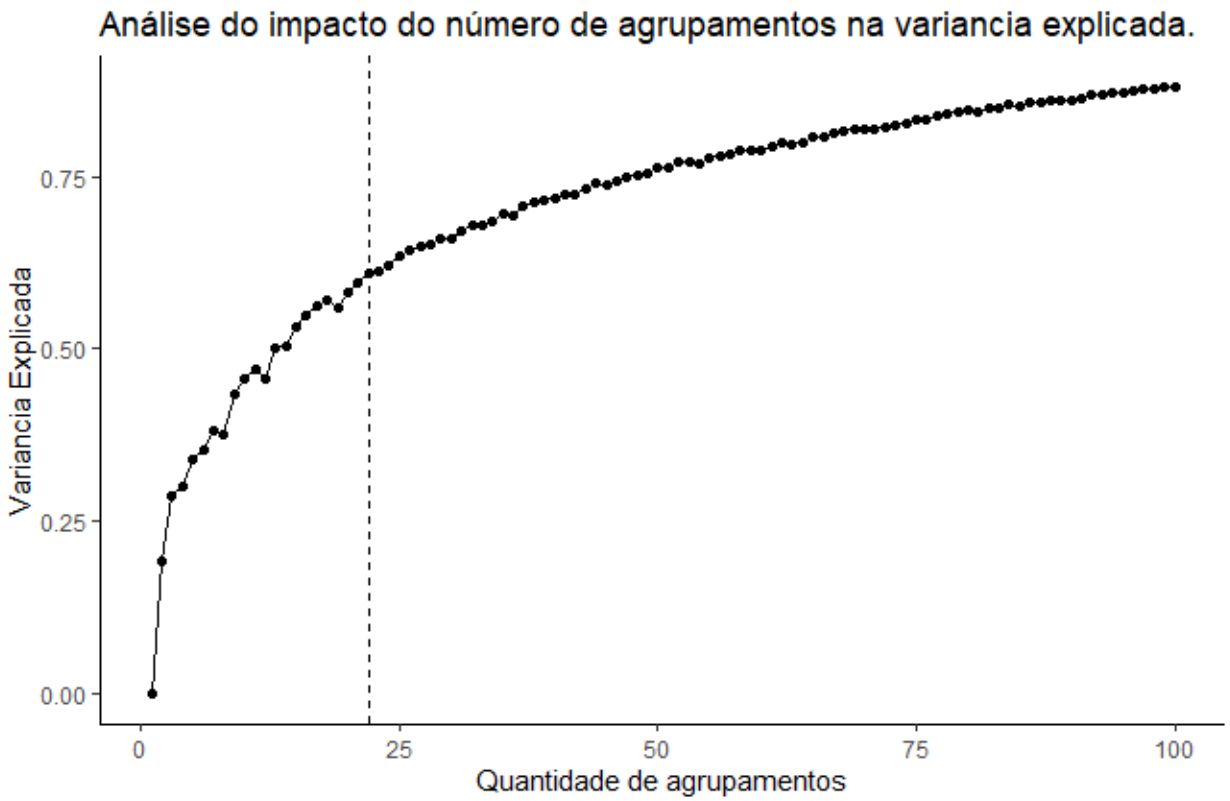


Figura 9: Seleção do número de agrupamentos mais representativo dos dados. Fonte: Figura do Autor.

Após a validação foi observado que o ganho de variância explicada deixa de ser tão expressivo após a marca dos 22 grupos. E assim o modelo final foi calculado utilizando 22 grupos.

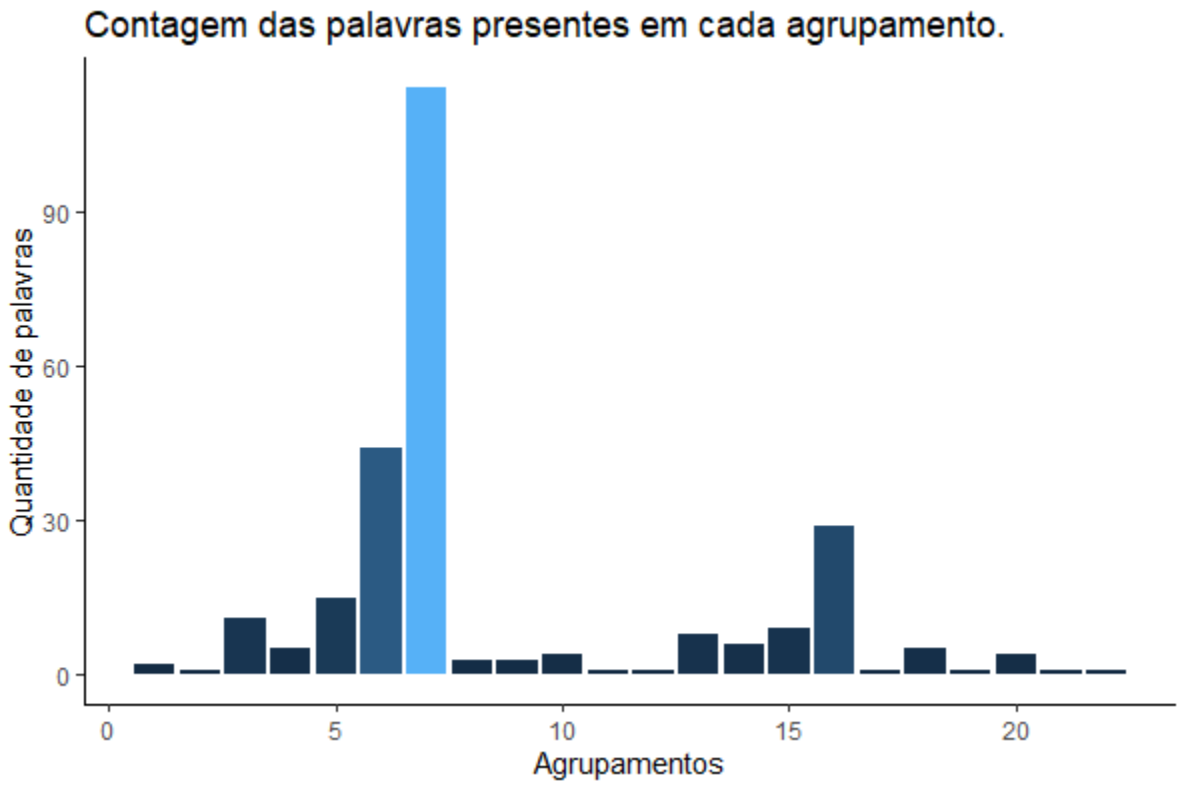


Figura 10: Distribuição da quantidade de palavras nos 22 grupos. Fonte: Figura do Autor.

Na figura acima é possível observar que os grupos não apresentam uma distribuição uniforme das palavras dentro deles, sendo que o grupo 7 apresenta a maioria das palavras. Por outro lado, vários grupos ficaram com apenas uma palavra representando este grupo.

Com um número tão baixo de palavras em certos grupos e a maioria contida no grupo 7, mas também nos grupos 6 e 16, as informações destes resultados apresentam baixas chances de apresentar alguma associação nova sobre os temas discutidos nos artigos científicos.

4.3. LDA

4.3.1. Validação Cruzada

A validação cruzada é uma técnica de avaliação do modelo antes de ser empregado procurando indicar quais parâmetros são mais apropriados para a análise a ser conduzida. Neste trabalho, a validação cruzada foi feita para investigar o impacto do número de tópicos na perplexidade observada no conjunto de dados, procurando equilibrar a capacidade de generalização do modelo e o tempo envolvido nos processos de treinamento e avaliação do modelo.

Foram testados: 5, 10, 25, 30, 40, 50, 60, 75 e 100 tópicos divididos em 5 partições iguais. Diferente da técnica anterior, a validação cruzada para o LDA apresentou uma quantidade de tempo bastante considerável e por isso não foi possível avaliar a diminuição da perplexidade em relação a cada número de tópicos.

Para cada quantidade de tópicos avaliada, foram criados 5 modelos baseados em diferentes combinações das partições que eram do grupo de treinamento e avaliação. As divisões sempre apresentavam 4 grupos de treinamento e 1 de avaliação, sendo que apenas a partir deste último foram calculadas as medidas de perplexidades.

Dessa forma, para cada quantidade de tópicos avaliada, temos 5 perplexidades diferentes e ao calcularmos a perplexidade média, podemos observar com maior acurácia a capacidade de generalização do modelo.

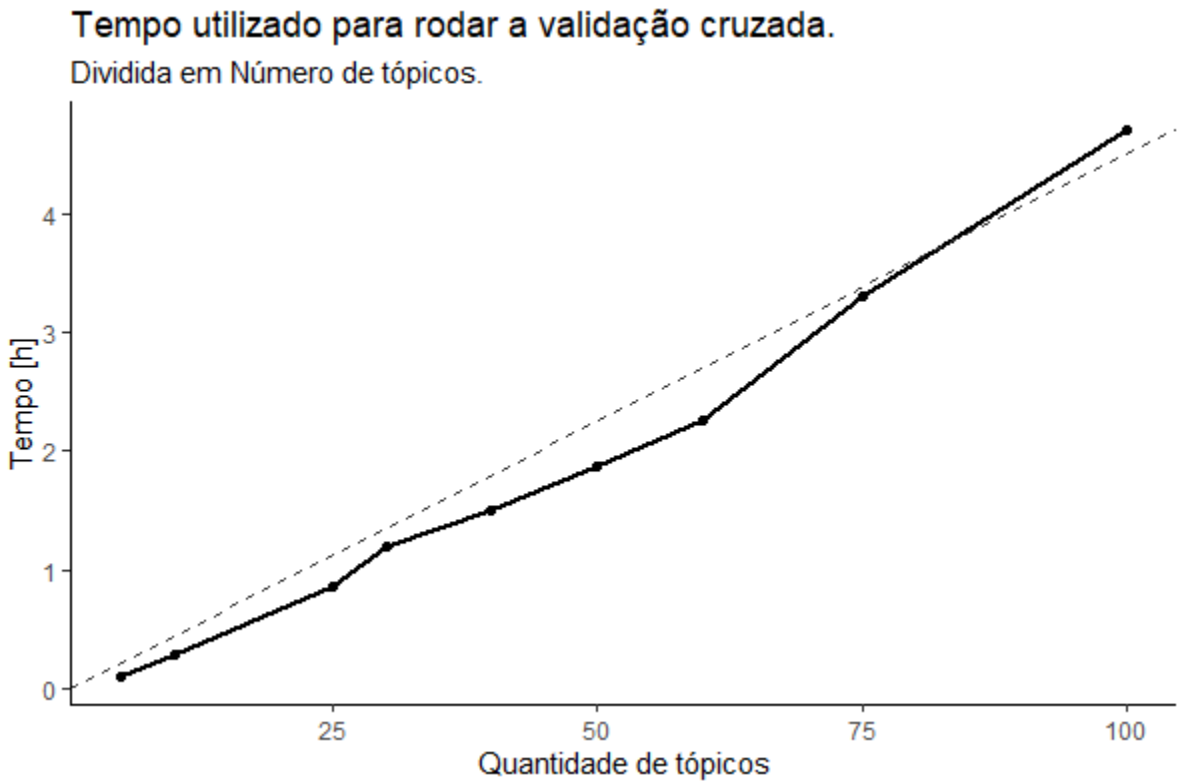


Figura 11: Tempo necessário para realizar a validação cruzada em função da quantidade de tópicos utilizada. Fonte: Figura do autor.

Diferente da técnica anterior que demorou alguns segundos em seu tempo de execução para avaliar os 100 modelos durante sua validação, para a técnica de LDA foi observada uma relação linear no tempo em horas necessário para avaliar cada número de tópicos, tornando a técnica muito custosa em relação a possibilidade de testes. Por outro lado, a validação cruzada adicionou maior robustez a análise graças às avaliações replicadas.

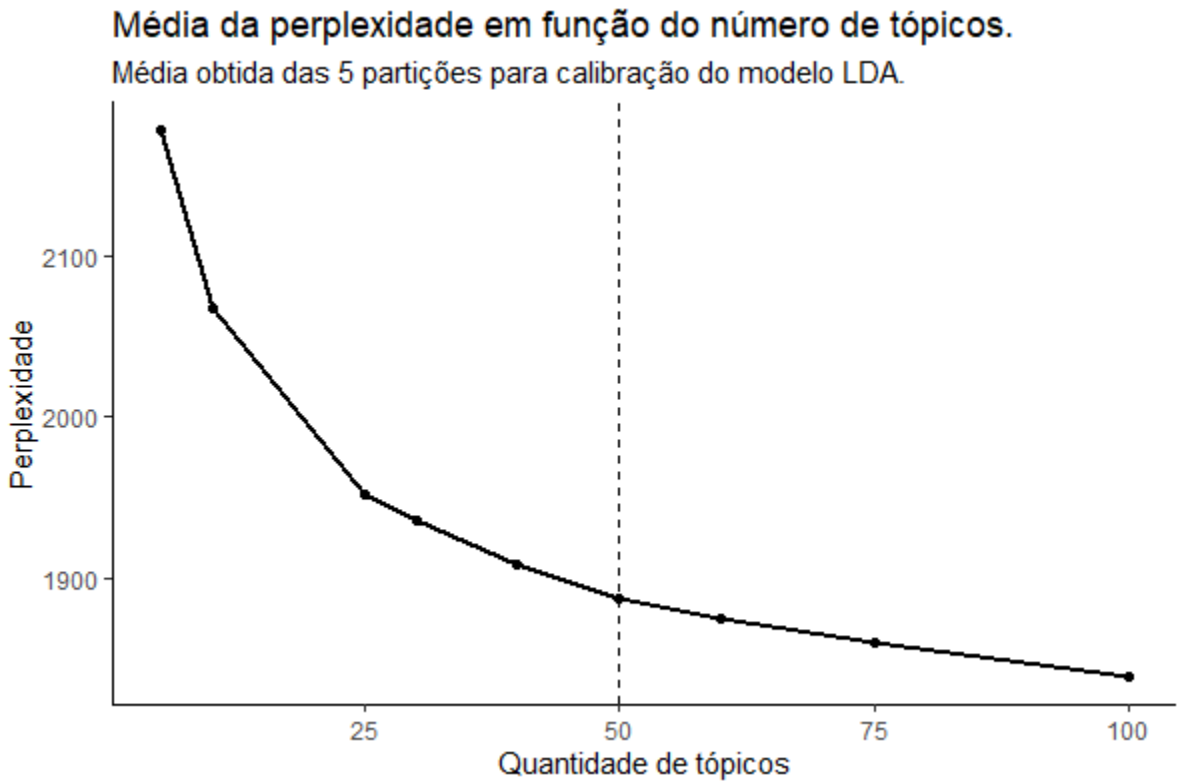


Figura 12: Visualização gráfica do número de tópicos na perplexidade média. Conforme a perplexidade média diminui, maior a capacidade do modelo de prever as palavras dentro de um dado tópico. Fonte: Figura do autor.

A partir das curvas geradas pelas simulações de Monte Carlo, podemos ver que em 50 tópicos observamos uma mudança na taxa de queda da perplexidade média. A perplexidade média continua a cair conforme mais tópicos ou agrupamentos são gerados, porém notamos uma diminuição na taxa de queda da perplexidade média após 50 tópicos indicando que esta é uma quantidade satisfatória para análise desse conjunto de dados.

É importante ressaltar que a perplexidade poderia ser reduzida em maior quantidade se algumas palavras menos frequentes fossem retiradas da matriz de termos. No entanto foi escolhida a opção de mantê-las, pois eram palavras específicas como o nome de antidepressivos e apesar de serem palavras “improváveis” de serem selecionadas pelo modelo para compor o texto do tópico a sua presença agregou valor a investigação do sentido dos tópicos latentes.

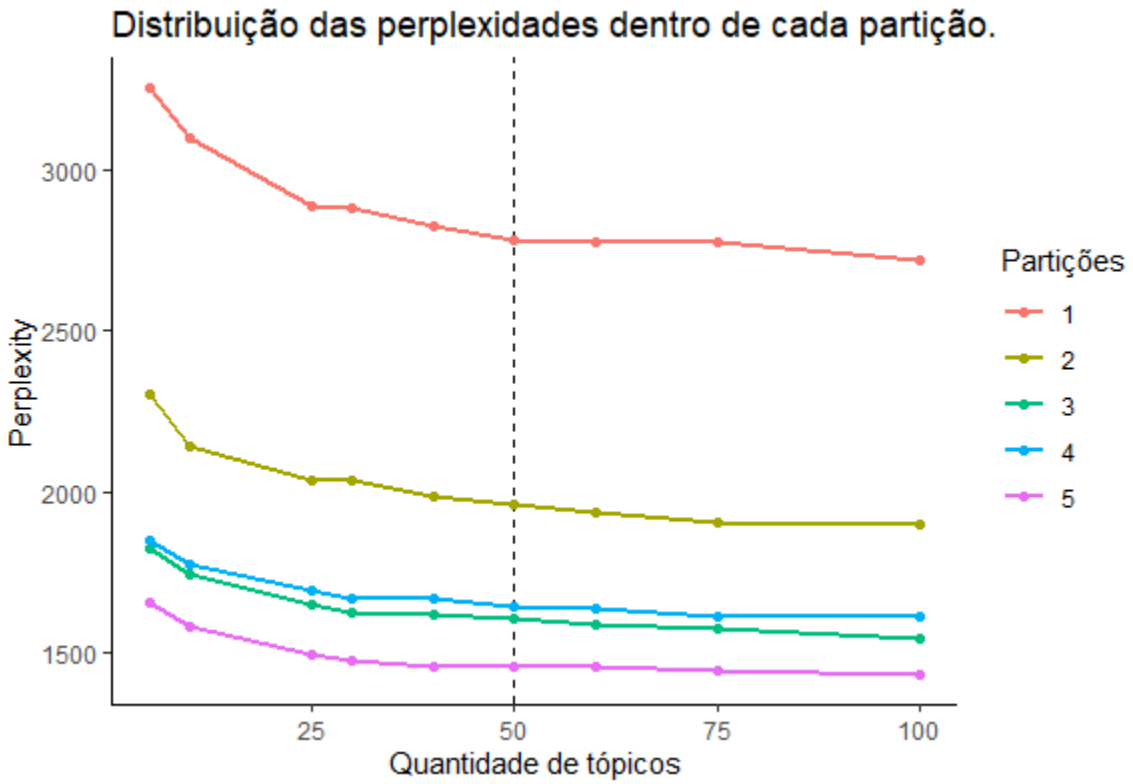


Figura 13: Perplexidade calculada em cada quantidade de tópicos analisada dividida por partições. Fonte: Figura do autor.

Também podemos observar que dentro das diferentes partições um comportamento geral é encontrado. Cada partição apresenta uma perplexidade diferente das outras, variação oriunda dos próprios dados, mas todos já estão tendendo a uma perplexidade “estável” por volta de 50 tópicos.

4.3.2. Distribuição dos artigos em cada tópico

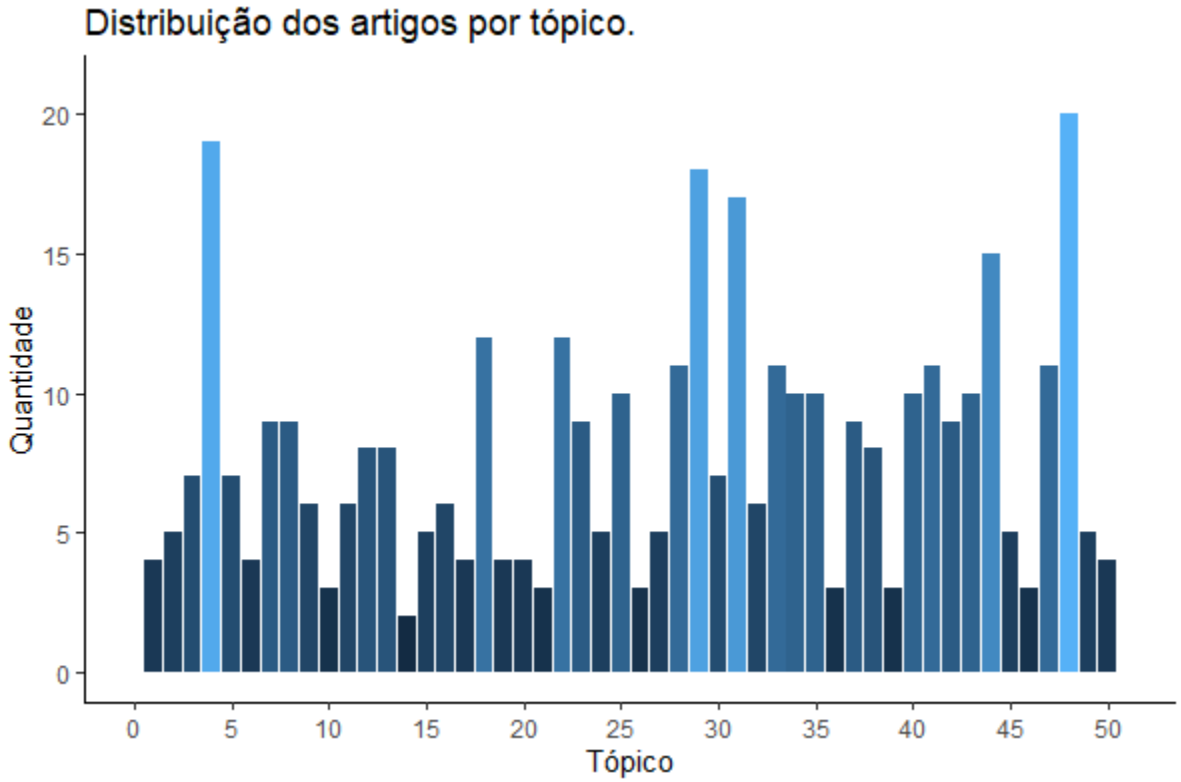


Figura 14: Frequência absoluta de artigos por tópicos. Fonte: Figura do autor.

Na figura acima podemos observar a distribuição dos artigos científicos em cada um dos 50 tópicos gerados. Apesar da técnica de LDA devolver uma probabilidade associada a cada tópico para os documentos, para essa representação gráfica o tópico com maior probabilidade foi definido como o tópico daquele documento.

Notamos que como na técnica anterior não existe homogeneidade entre os tópicos, porém a discrepância se tornou bem menor entre os tópicos. Esta observação se dá devido à natureza dos artigos usados para a avaliação, apesar de serem todos artigos científicos sobre antidepressivos, alguns temas são mais prevalentes e tem maior número de publicações disponíveis.

4.3.3. Descrição dos tópicos

Após o processo de agrupamentos realizado pela técnica de LDA, os tópicos formados foram avaliados a partir das palavras com maior valor de beta calculado e os títulos dos artigos. Quando o tópico ainda não era claro, o conteúdo da discussão e da conclusão dos artigos também foram usados para criar uma descrição geral do conteúdo de cada um dos tópicos. Em alguns casos, apesar de todas essas informações, nenhum tópico ficava evidente em relação ao corpo de artigos agrupado. Nestes casos, o tópico foi classificado como “Sem tópico claro.”.

Abaixo estão listadas as descrições dos 50 tópicos gerados pela técnica de LDA:

1. Desenvolvimento de complexos para melhor performance de antidepressivos.
2. Efeitos disruptivos da nicotina e de opioides causando aumento de sintomas depressivos.
3. Melhoras morfológicas por antidepressivos e terapia eletroconvulsiva em pacientes com depressão.
4. Efeito de fitofármacos, metformina e imipramina na diminuição dos sintomas de depressão.
5. Efeitos de antidepressivos baseados em modelos animais aquáticos e impactos na fauna aquática de antidepressivos.
6. Impactos da depressão, melancolia, outras neuropatologias e antidepressivos na memória.
7. Extratos naturais, estatinas e coadjuvantes atuando sobre receptores de agmatina.
8. Ansiolíticos e antidepressivos com efeitos sobre dor neuropática e imuno modulação.
9. Estudos epidemiológicos com foco na aderência ao tratamento.
10. Antidepressivos e seus efeitos em diferentes regiões cerebrais.
11. Química analítica e sensores químicos para antidepressivos.
12. Síntese e avaliação biológica de antidepressivos.
13. Avaliação de biomarcadores e efeito de vitaminas no tratamento da depressão.
14. Visão social da procura e efeitos de neuro fármacos.

15. Modelos computacionais na descoberta de novos fármacos e funcionamento de receptores.
16. Tratamento de comorbidades durante o tratamento da depressão.
17. Sem tópico claro.
18. Identificação e quantificação de antidepressivos em diferentes meios.
19. Sem tópico claro.
20. Influência da cetamina na libido de indivíduos com depressão e da serotonina nos sentidos.
21. Fisiopatologia da depressão e o impacto do processo inflamatório no cérebro.
22. Regiões e conexões cerebrais associadas ao desenvolvimento de depressão.
23. Redução de neuro inflamação e redução de danos por estresse oxidativo.
24. Moléculas coadjuvantes no tratamento de depressão resistente a tratamento.
25. Efeitos do estresse nas futuras gerações.
26. Sem tópico claro.
27. Neurônios dopaminérgicos na depressão.
28. Estudos de coorte avaliando comportamento suicida.
29. Tratamento por estimulação magnética transcranial.
30. Poluição aquática causada por antidepressivos.
31. Intervenções farmacológicas e não farmacológicas no tratamento da depressão.
32. Estudos in vitro e in vivo de farmacodinâmica e farmacocinética.
33. Uso de hipnóticos e sedativos em especial em pacientes idosos com demência.
34. Degradação e farmacocinética de diferentes antidepressivos.
35. Avaliação da depressão em diferentes fases da vida feminina.
36. Avaliação da mudança de hábitos em camundongos após tratamento com antidepressivos.
37. Avaliação do tratamento com cetamina na depressão.
38. Relações entre a microbiota intestinal e depressão.
39. Sem tópico claro.
40. Avaliação da atividade antidepressiva e ansiolítica da cetamina e moléculas derivadas.
41. Avaliando o impacto da redução de neuro inflamação em depressão e ansiedade.

42. Modelos de previsão de depressão e resposta a antidepressivos.
43. Estudos sobre neuro toxinas.
44. Estudos de coorte avaliando diversas patologias.
45. Novas propostas de fitofármacos para o tratamento de ansiedade e depressão.
46. Impacto de fatores ambientais, como período de sono reduzido e atividade física, em sintomas da depressão.
47. Avaliação do impacto de traumas e negligencia na infância de pacientes com depressão.
48. Testes clínicos avaliando a eficácia de novos fármacos e novos tratamentos para depressão.
49. Detecção e impactos de eco toxinas, em especial fármacos.
50. Diferentes usos de antidepressivos, ex.: antibiótico ou estimulando reparo de tecido ósseo.

4.3.4. Cetamina

Dentro da nuvem de palavras apresentada acima, podemos notar que alguns termos estão bastante presentes nos documentos analisados. Em especial, um fármaco apareceu vezes o bastante para estar entre os top 50 termos: a cetamina.

A cetamina atua de forma distinta dos inibidores de recaptação de serotonina, pois ela inibe os receptores NMDA, uma classe de receptores glutamatérgicos (BERMAN et al., 2000). Este é um fármaco que foi fundamental para o desenvolvimento da hipótese neurotrópica da depressão e que demonstrou potencial para o tratamento de depressão resistente ao tratamento é a cetamina (ZARATE; BRUTSCHE; CHARNEY, 2006). Diferente dos outros antidepressivos clássicos, a cetamina apresentou melhora de sintomas em pacientes entre 2 a 4 horas após a primeira infusão e seus efeitos duraram até 7 dias após a administração, mesmo em pacientes resistentes aos tratamentos anteriores (BERMAN et al., 2000; ZARATE; BRUTSCHE; CHARNEY, 2006). Após inibir os receptores de NMDA, é observado um aumento considerável de glutamato no cortéx pré-frontal e há indícios de que este efeito estimule a plasticidade sináptica através da liberação de BDNF (DUMAN et al., 2016). Além disso, a cetamina também estimula a

secreção de VEGF no hipocampo (CHOI *et al.*, 2016) e ambos os fatores apresentam sinergia para os efeitos antidepressivos da cetamina (DEYAMA; DUMAN, 2020).

Por ser um fármaco com tamanha relevância nos documentos analisados, dentro do modelo gerado, o parâmetro gamma descreve a magnitude do quanto um documento pertence a um tópico. Quando distribuímos anteriormente os documentos, o gráfico n+8 foi criado utilizando o maior valor de gamma para cada um dos documentos. Porém, em alguns casos o valor de gamma para dois ou mais tópicos ficam próximos, indicando que o documento tem similaridade com mais de um tema.

Dos 3 temas que foram descritos com cetamina como um elemento central: influência da cetamina na libido de indivíduos com depressão e da serotonina nos sentidos, avaliação do tratamento com cetamina na depressão e avaliação da atividade antidepressiva e ansiolítica da cetamina e moléculas derivadas, estão presentes 23 documentos. Após a avaliação de quais destes documentos apresentavam valores de gamma para outros tópicos que fossem de magnitude similar ao do tópico principal, sobraram 8 documentos.

Tabela 1 – Documentos com valores de gamma próximos.

Título	Tópico 1	Gamma 1	Tópico 2	Gamma 2
Association of parental depression with offspring attention deficit hyperactivity disorder and autism spectrum disorder: A nationwide birth cohort study	37	0.41	35	0.38
Neurocognitive performance of repeated versus single intravenous subanesthetic ketamine in treatment resistant depression	37	0.42	3	0.39
Ketamine Enhances Visual Sensory Evoked Potential Long-term Potentiation in Patients With Major Depressive Disorder	37	0.35	22	0.31
Sub-anesthetic and anesthetic ketamine produce different long-lasting behavioral phenotypes (24 h post-treatment) via inducing different brain derived neurotrophic factor (BDNF) expression level in the hippocampus	40	0.4	7	0.32
Antidepressant effects of ketamine on depression-related phenotypes and dopamine dysfunction in rodent models of stress	40	0.43	25	0.28
An old but still burning problem: Inter-rater reliability in clinical trials with antidepressant medication	37	0.58	48	0.42
Subanesthetic ketamine exerts antidepressant-like effects in adult rats exposed to juvenile stress	40	0.49	25	0.24
Long-term outcome in outpatients with depression treated with acute and maintenance intravenous ketamine: A retrospective chart review	37	0.56	28	0.23

Fonte: Tabela do autor.

Considerando a Tabela 1, observamos casos de associações entre os tópicos sobre cetamina e os tópicos sobre: estudos clínicos, efeito do estresse em diferentes momentos da vida dos pacientes e também na geração seguinte, em mudanças morfológicas no cérebro e em fármacos que estão sendo estudados por apresentarem potencial como antidepressivos.

Frente a essas informações, um seguimento de estudos interessante e que pode trazer novas informações relevantes é a avaliação do impacto da cetamina no tratamento da depressão em mães e nas próximas gerações, visando investigar possíveis mudanças morfológicas no sistema nervoso central, em especial no córtex pré-frontal e no hipocampo.

5. Conclusão:

Este trabalho se fundamentou em técnicas matemáticas para agrupar e analisar documentos textuais sobre antidepressivos publicados entre 2020 e 2021. Através dos documentos foi possível descobrir os principais termos envolvidos em cada artigo científico, gerar um agrupamento baseado apenas nas palavras envolvidas e compara-lo a um agrupamento levando em consideração um modelo misto de probabilidades.

Para o modelo de K-médias, encontramos um ponto ótimo em 22 grupos, porém a distribuição dos documentos se apresentou de uma forma consideravelmente assimétrica e não foi possível gerar grupos claros a partir desse agrupamento. Já para o modelo de LDA, encontramos um ponto ótimo em 50 tópicos e, a partir dessa avaliação, os 50 tópicos foram descritos a partir das principais palavras, dos títulos dos documentos e também das discussões e conclusões dos documentos.

Além disso, com foco especial nos tópicos associados a cetamina, procuramos os artigos científicos que apresentassem valores de parâmetro gamma para outros tópicos que fossem relevantes e encontramos indícios de que o estudo da influência da cetamina em mães e suas próximas gerações trarão informações relevantes para o tratamento da depressão nesses grupos.

6. Referências Bibliográficas:

AIZAWA, Akiko**. An information-theoretic perspective of tf–idf measures**, Information Processing & Management, v. 39, n. 1, p. 45-65, 2003.

ANDERBERG, Michael. **Cluster Analysis for Applications**. Academic Press, 1973.

ASMUSSEN, Claus Boye; MØLLER, Charles. **Smart literature review**: a practical topic modelling approach to exploratory literature review. Journal of Big Data, v. 6, n.193, 2019.

Associação Americana de Psiquiatria. **What Is Depression?** Disponível em: <https://www.psychiatry.org/patients-families/depression/what-is-depression>. Acessado em: 03/10/2021

BERMAN, Robert M.; CAPPIELLO, Angela; ANAND, Amit; OREN, Dan A.; HENINGER, George R.; CHARNEY, Dennis S.; KRYSTAL, John H. **Antidepressant Effects of Ketamine in Depressed Patients**, Society of Biological Psychiatry, v. 47, p. 351-354, 2000.

BERRAR, Daniel. **Cross Validation**. Encyclopedia of Bioinformatics and Computational Biology, Academic Press, p. 542-545, 2019.

BLEI, David; NG, Andrew; JORDAN, Michael. **Latent Dirichlet Allocation**. Journal of Machine Learning Research, v. 3, p. 993-1022, 2003.

CHENG, Lionel et al. **Discerning tumor status from unstructured MRI reports: completeness of information in existing reports and utility of automated natural language processing**. Journal of Digital Imaging, v. 23, n. , p. 119-132, 2010.

CHOI, Miyeon; LEE, Seung H.; HO, Lee C.; SON, Hyeon. **Hippocampal VEGF is necessary for antidepressant-like behaviors but not sufficient for antidepressant-like effects of ketamine in rats**, Molecular Basis of Disease, v. 1862, p. 1247-1254, 2016.

DEYAMA, Satoshi; DUMAN, Ronald S. **Neurotrophic mechanisms underlying the rapid and sustained antidepressant actions of ketamine**, v. 188, 2020.

DUMAN, Ronald S.; AGHAJANIAN, George K.; SANACORA, Gerard; KRYSTAL, John H. **Synaptic plasticity and depression**: new insights from stress and rapid-acting antidepressants, Nature Medicine, v. 22, p. 238-249, 2016.

DUMAN, Ronald S.; LI, Nanxin. **A neurotrophic hypothesis of depression**: role of synaptogenesis in the actions of NMDA receptor antagonists, Philosophical Transactions of The Royal Society, v. 367, p. 2475-2484, 2012.

EISENSTEIN, Jacob. **Natural Language Processing**. Primeira edição. Editora do MIT, publicado em 2018.

FAWCETT, Tom. **Data Science for Business**: What You Need to Know about Data Mining and Data-Analytic Thinking, O'Reilly Media Inc., 2013.

FREIS, Edward D. **Mental Depression in Hypertensive Patients Treated for Long Periods with Large Doses of Reserpine**, The New England Journal of Medicine, v. 251, p. 1006-1008, 1954.

GRIFFITHS, Thomas; STEYVERS, Mark. **Finding Scientific Topics**. Proceedings of the National Academy of Sciences of the United States of America, v. 101, p. 5228-5235, 2004.

JAIN, Anil Kumar; MURTY, Narasimha; FLYNN, Patrick. **Data Clustering**: A Review. Association for Computing Machinery, ACM Comput. Surv., v. 31, p. 264-323, 1999.

KOBAYASHI, Hayato. **Perplexity on Reduced Corpora**. Association for Computational Linguistics, p. 797-806, 2014.

KRISHNAN, Vaishnav; NESTLER, Eric J. **The molecular neurobiology of depression**, Nature, v. 455, p. 894-902, 2008.

NUTT, David J.; FORSHALL, Sam; BELL, Caroline; RICH, Ann; SANDFORD, John; NASH, Jon; ARGYROPOULOS, Spilios. **Mechanisms of action of selective serotonin reuptake inhibitors in the treatment of psychiatric disorders**, European Neuropsychopharmacology, v.9, p. 81-86, 1999.

OMS. **Depression**. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/depression>. Acesso em: 03/10/2021.

ROBERTSON, Stephen. **Understanding Inverse Document Frequency**:On theoretical arguments for IDF, Journal of Documentation, v. 60, n. 5 p. 503-520, 2004.

SAS. **Evolution of machine learning**. Disponivel em:

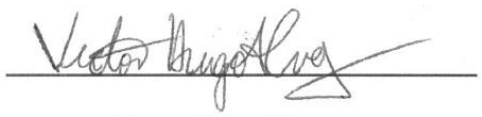
< https://www.sas.com/en_us/insights/analytics/machine-learning.html >. Acesso em: 01/10/2021.

TRIVEDI, Madhukar H.; RUSH, John A.; WISNIEWSKI, Stephen R.; NIERENBERG, Andrew A.; WARDEN, Diane; RITZ, Louise; et al. **Evaluation of Outcomes With Citalopram for Depression Using Measurement-Based Care in STAR*D**: Implications for Clinical Practice, The American Journal of Psychiatry, v. 163, p. 28-40, 2006.

ZARATE, Carlos A.; BRUTSCHE, Nancy; CHARNEY, Dennis. **A Randomized Trial of na N-Methyl-D-Aspartate Antagonist in Treatment-Resistant Major Depression**, General Psychiatry, v. 63, p. 856-864, 2006.

Web of Science, **Antidepressants**. Disponível em:

<https://www.webofscience.com/wos/woscc/summary/19accd99-7c57-43cb-ba1a-6cdaca106af8-11510e4c/relevance/1>. Acesso em: 05/11/2021.



Victor Hugo Alves

05/11/2021

Prof. Dr. Gabriel Lima Barros de Araújo

05/11/2021