

Luiz Felipe Matos de Melo

**Identificando Temas Relacionados à Depressão
no *Reddit* Usando Modelagem de Tópicos**

Niterói, RJ, Brasil

2021

Luiz Felipe Matos de Melo

Identificando Temas Relacionados à Depressão no *Reddit* Usando Modelagem de Tópicos

Trabalho submetido ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal Fluminense

Instituto de Computação

Departamento de Ciência da Computação

Orientadora: Profa. Aline Marins Paes Carvalho

Coorientador: Paulo Roberto Mann Marques Júnior

Niterói, RJ, Brasil

2021

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

M528i Melo, Luiz Felipe Matos de
Identificando Temas Relacionados à Depressão no Reddit
Usando Modelagem de Tópicos / Luiz Felipe Matos de Melo ;
Aline Marins Paes Carvalho, orientadora ; Paulo Roberto Mann
Marques Júnior, coorientador. Niterói, 2021.
98 f. : il.

Trabalho de Conclusão de Curso (Graduação em Ciência da
Computação)-Universidade Federal Fluminense, Instituto de
Computação, Niterói, 2021.

1. Inteligência artificial. 2. Aprendizado de máquina. 3.
Processamento de linguagem natural (Computação). 4. Rede
social. 5. Produção intelectual. I. Carvalho, Aline Marins
Paes, orientadora. II. Marques Júnior, Paulo Roberto Mann,
coorientador. III. Universidade Federal Fluminense. Instituto
de Computação. IV. Título.

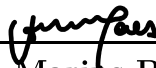
CDD -

Luiz Felipe Matos de Melo

Identificando Temas Relacionados à Depressão no *Reddit* Usando Modelagem de Tópicos

Trabalho submetido ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.


Trabalho aprovado. Niterói, RJ, Brasil, 06 de Maio de 2021:



Profa. Aline Marins Paes Carvalho
(D.Sc.)
Orientadora - UFF



Paulo Roberto Mann Marques Júnior
(M.Sc.)
Coorientador - UFF



Daniel Cardoso Moraes de Oliveira
(D.Sc.)
Convidado 1 - UFF



Daniela Quitete de Campos Vianna
(Ph.D.)
Convidada 2 - UFF

Niterói, RJ, Brasil

2021

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais, Noemi e Luiz, por terem me dado o suporte e auxílio necessários para que eu conseguisse prosseguir nos estudos. Diante de todas as dificuldades impostas pela vida, vocês fizeram o possível e o impossível para proporcionar a mim algo que vocês não tiveram a oportunidade de usufruir em sua totalidade durante a juventude: o acesso ao ensino superior. Tenho a certeza hoje que minhas conquistas não poderiam ter ocorrido sem que minha mãe, nos primórdios da minha infância, tivesse ensinado para mim a beleza da leitura, e sem que meu pai tivesse feito o máximo para me proporcionar um computador, mesmo que ele não entendesse bem o que a tal máquina fazia e que a mesma conferisse um débito proibitivo nas nossas (bastante) modestas finanças. Serei eternamente grato por tudo isso.

Cito também minha namorada Lídia, sempre disposta a me apoiar nos momentos de maior estresse com a graduação e em compartilhar os momentos mais simples do dia-a-dia. Sua presença e bom humor são sempre de especial importância para mim. Ainda, cito aqui meus amigos Vitor e Pedro, que conheci durante a graduação e com os quais passei muitas noites estudando e conversando sobre os mais diversos assuntos. Em todo esse tempo, demos muitas risadas juntos, mas também passamos por alguns bocados relativos às dificuldades típicas da graduação e do início da vida adulta.

Ainda, devo mencionar aqui meus professores do ensino fundamental e médio, que mostraram que eu poderia ir mais longe independentemente da minha origem. Eu já não lembro os nomes de muitos de vocês, mas todos os incentivos que me deram contribuíram para que me tornasse hoje a pessoa que sou. Obrigado por não desistirem de proporcionar uma boa educação, mesmo com o constante descaso do Estado para com os setores básicos de ensino público. É devido a pessoas como vocês, que gente como eu tem a possibilidade de progredir na vida.

E falando em professores, isso me proporciona a brecha para agradecer especialmente a meus orientadores neste trabalho, Aline Paes e Paulo Mann. A orientação proporcionada por vocês aprimorou de forma imensa a realização desta monografia. Obrigado por estarem dispostos a solucionar as dúvidas e sugerir caminhos alternativos a cada problema com que me deparei no curso do estudo. Sobretudo, sou muito grato pela compreensão e paciência com minha rotina dupla de estudante e trabalhador, que por vezes impediu que eu agisse com maior celeridade para o avanço de algumas etapas do estudo. Devo citar, ainda, o esforço empreendido por vocês para proporcionar uma orientação com tanta qualidade em um período especialmente difícil como o atual, mantendo reuniões frequentes e acompanhamento constante. Tenho orgulho de ter sido orientado por dois excelentes

pesquisadores como vocês, e aprendi bastante com ambos.

Devo citar também o professor Daniel Cardoso e a colega Daniela Vianna, que prontamente aceitaram participar da banca avaliadora deste trabalho. Espero aprimorar meu conhecimento a partir das críticas e sugestões que advenham de sua leitura e análise da presente monografia.

Por fim, não poderia deixar de agradecer à Universidade Federal Fluminense, e em especial ao Instituto de Computação, por proporcionarem um ensino de excelência em um momento como o atual. É um privilégio tremendo fazer parte da comunidade de estudantes da UFF, e sempre irei lembrar com carinho do período da graduação. A UFF me apoiou desde o início, provendo auxílio-permanência assim que iniciei meus estudos. Dentro dela, tive a oportunidade de conhecer excelentes pessoas, grandes pesquisadores e aprender como o conhecimento científico é construído e aprimorado – tanto durante a iniciação científica quanto durante a elaboração desta monografia. O trabalho realizado pela UFF, e pelo ensino superior público brasileiro como um todo, é um trabalho que deve ser reconhecido e apoiado incondicionalmente.

A importância de uma instituição como a UFF torna-se ainda mais relevante em um momento conturbado como o vivido nesse Brasil dos anos 2020. Hoje, como diria Carl Sagan, os demônios voltaram a assombrar a racionalidade com força total, promovendo a desinformação, a pseudo-ciência e curas fáceis para problemas complexos, tentando a todo custo apagar a única vela que pode nos guiar atualmente: a ciência. Contudo, da mesma forma que podemos acender uma vela esfregando pedras, podemos fazer ciência apesar do negacionismo. Diante do discutido, eu não posso sentir nada além do mais amplo orgulho em estar concluindo um curso de graduação na UFF, e serei eternamente grato pelas oportunidades que a instituição proveu a mim e a muitos outros estudantes como eu.

“[...] Repito que é melhor a verdade por dura que seja que uma fantasia consoladora. E, na hora da verdade, os fatos revistam ser mais reconfortantes que a fantasia.”
(Carl Sagan, em “O mundo assombrado pelos demônios”)

Resumo

Nas últimas décadas, os níveis de prevalência do transtorno depressivo têm se elevado em todo o planeta. As consequências debilitantes da depressão impactam negativamente os mais variados âmbitos da vida de um acometido pelo transtorno. Um problema de saúde pública mundial, a depressão passou a ser o foco de diversos estudos destinados a identificar possíveis casos na população geral e prover tratamento adequado aos indivíduos. Por outro lado, nos últimos anos, as redes sociais tornaram-se ferramentas de comunicação predominantes e inerentes à vida moderna. Contudo, pouco se sabe sobre seus impactos na mente humana. Alguns estudos buscam mostrar ligações entre o uso excessivo das redes e o surgimento de sintomas de doenças mentais, incluindo depressão, enquanto outros mostram o potencial das redes no suporte a usuários com tais problemas. Dado que a depressão não se restringe às barreiras continentais ou idiomáticas, estudar o transtorno no âmbito de diferentes linguagens pode ser uma abordagem útil para ampliação de seu entendimento. No presente estudo, o objetivo é aplicar técnicas de processamento de linguagem natural como a modelagem de tópicos em fóruns de discussão relacionados à depressão na rede social *Reddit*. Desta forma, deseja-se levantar os temas relevantes para usuários de tais comunidades no contexto da depressão, nas línguas portuguesa e inglesa. Sendo assim, será possível determinar: 1) a aplicabilidade das técnicas de processamento de linguagem natural para levantamento de temas associados à depressão no *Reddit*; e 2) as possíveis semelhanças e diferenças entre os tópicos relevantes para a depressão, para falantes de ambos os idiomas.

Palavras-chaves: depressão, redes sociais, processamento de linguagem natural, modelagem de tópicos.

Abstract

In the last decades, the prevalence of the major depressive disorder has risen on a global scale. The debilitating consequences of depression negatively impact multiple areas of a depressed person's life. A worldwide public health problem, depression, became the focus of many studies to identify possible cases in the general population and to provide them the adequate treatment. On the other hand, in the last years, social networks have become predominant communication tools and inherent to modern life. However, little is known about their impacts on the human mind. Some studies aim to show the connections between excessive use of social networks and the emergence of mental disease symptoms, including depression, while others have shown the social networks' potential to provide support for users with mental health problems. Given that depression is not restricted by continental or language barriers, studying the disorder in the context of different languages can be a valuable approach for better understanding it. The present study aims at applying natural language processing techniques, particularly, topic modeling in depression-related discussions on Reddit forums. Thus, the aim is to find relevant topics related to depression for users of these communities, in both Portuguese and English languages. With that, we expect to determine: 1) the applicability of natural language processing techniques to find themes associated with depression in Reddit, and 2) the possible similarities and differences between relevant topics for depression for both Portuguese and English-speaking users.

Keywords: depression, social networks, natural language processing, topic modeling.

“[...] Repito que é melhor a verdade por dura que seja que uma fantasia consoladora. E, na hora da verdade, os fatos revistam ser mais reconfortantes que a fantasia.”
(Carl Sagan, em “O mundo assombrado pelos demônios”)

Lista de ilustrações

Figura 1 – Evolução do número de usuários ativos mensais em diversas redes sociais no período 2004-2019. Imagem obtida de (DATA, 2020).	2
Figura 2 – Codificação BOW dos termos de um documento e do próprio documento.	9
Figura 3 – Arquitetura de uma rede neural artificial para classificação de imagens. Utiliza fotografia obtida de (MARCHAL, 2005).	12
Figura 4 – Codificação <i>word2vec</i> dos termos de um documento.	13
Figura 5 – Arquiteturas dos modelos CBOW e <i>continuous skipgram</i> , obtidas de (RONG, 2014).	14
Figura 6 – Arquitetura do BERT em alto-nível, obtida de (DEVLIN et al., 2018).	16
Figura 7 – Codificação BERT dos termos de um documento.	17
Figura 8 – Processo generativo dos modelos de tópicos estudados. Trecho de <i>O Processo</i> de Franz Kafka usado para exemplificar o documento construído (KAFKA, 2019).	20
Figura 9 – Fases de construção do estudo realizado.	32
Figura 10 – Histograma de palavras da faixa de frequência de interesse a ser filtrada no <i>corpus</i> em português.	39
Figura 11 – Histograma de palavras da faixa de frequência de interesse a ser filtrada no <i>corpus</i> em inglês.	40
Figura 12 – Variação da coerência NPMI à medida que o número de tópicos é incrementado nos modelos em português.	48
Figura 13 – T_0 do modelo LDA	50
Figura 14 – T_1 do modelo LDA	50
Figura 15 – T_2 do modelo LDA	50
Figura 16 – T_3 do modelo LDA	50
Figura 17 – T_4 do modelo LDA	50
Figura 18 – Cinco categorias lexicais predominantes nos tópicos do modelo LDA com $K = 5$ treinado em português, segundo o LIWC.	52
Figura 19 – Variação da coerência NPMI à medida que o número de tópicos é incrementado nos modelos em inglês.	55
Figura 20 – T_0 do modelo ETM	59
Figura 21 – T_1 do modelo ETM	59
Figura 22 – T_2 do modelo ETM	59
Figura 23 – T_3 do modelo ETM	59
Figura 24 – T_4 do modelo ETM	59
Figura 25 – T_5 do modelo ETM	59
Figura 26 – T_6 do modelo ETM	59

Figura 27 – T_7 do modelo ETM	59
Figura 28 – T_8 do modelo ETM	61
Figura 29 – T_9 do modelo ETM	61
Figura 30 – T_{10} do modelo ETM	61
Figura 31 – T_{11} do modelo ETM	61
Figura 32 – T_{12} do modelo ETM	61
Figura 33 – T_{13} do modelo ETM	61
Figura 34 – T_{14} do modelo ETM	61
Figura 35 – T_{15} do modelo ETM	61
Figura 36 – T_{16} do modelo ETM	64
Figura 37 – T_{17} do modelo ETM	64
Figura 38 – T_{18} do modelo ETM	64
Figura 39 – T_{19} do modelo ETM	64
Figura 40 – T_{20} do modelo ETM	64
Figura 41 – T_{21} do modelo ETM	64
Figura 42 – T_{22} do modelo ETM	64
Figura 43 – T_{23} do modelo ETM	64
Figura 44 – T_{24} do modelo ETM	65
Figura 45 – T_{25} do modelo ETM	65
Figura 46 – T_{26} do modelo ETM	65
Figura 47 – T_{27} do modelo ETM	65
Figura 48 – Cinco categorias lexicais predominantes nos tópicos do modelo ETM com $K = 28$ treinado em inglês, segundo o <i>Empath</i>	66

Lista de tabelas

Tabela 1 – Informações sobre a coleta de submissões em ambos os idiomas.	35
Tabela 2 – Faixas de frequência em documentos e número de <i>tokens</i> únicos existentes para o <i>corpus</i> em português.	38
Tabela 3 – Faixas de frequência em documentos e número de <i>tokens</i> únicos existentes para o <i>corpus</i> em inglês.	39
Tabela 4 – Divisão de documentos para treino e validação em ambos os <i>corpora</i> estudados.	41
Tabela 5 – Modelos LDA em português por ordem decrescente de coerência.	47
Tabela 6 – Modelos CTM em português por ordem decrescente de coerência.	47
Tabela 7 – Modelos ETM em português por ordem decrescente de coerência.	47
Tabela 8 – Modelos em português por ordem decrescente de coerência.	47
Tabela 9 – Tópicos do modelo LDA com $K = 5$ em ordem decrescente de coerência.	49
Tabela 10 – Modelos LDA em inglês por ordem decrescente de coerência.	53
Tabela 11 – Modelos CTM em inglês por ordem decrescente de coerência.	53
Tabela 12 – Modelos ETM em inglês por ordem decrescente de coerência.	53
Tabela 13 – Modelos em inglês por ordem decrescente de coerência.	53
Tabela 14 – Tópicos do modelo ETM com $K = 28$ em ordem decrescente de coerência.	56

Lista de abreviaturas e siglas

OMS	Organização Mundial da Saúde
TDM	Transtorno Depressivo Maior
NLP	<i>Natural Language Processing</i>
LDA	<i>Latent Dirichlet Allocation</i>
CTM	<i>Contextualized Topic Model</i>
ETM	<i>Embedded Topic Model</i>
BOW	<i>Bag-of-words</i>
ANN	<i>Artificial Neural Networks</i>
CBOW	<i>Continuous Bag-of-words</i>
MLM	<i>Masked Language Model</i>
NSP	<i>Next Sentence Prediction</i>
VB	<i>Variational Bayes</i>
VAE	<i>Variational Autoencoder</i>
ELBO	<i>Evidence Lower Bound</i>
LIWC	<i>Linguistic Inquiry and Word Count</i>
POS	<i>Part-of-speech</i>
FD	<i>Frequência em Documentos</i>
NPMI	<i>Normalized Pointwise Mutual Information</i>
CPU	<i>Central Processing Unit</i>

Sumário

1	INTRODUÇÃO	1
1.1	Objetivos	3
1.2	Metodologia	4
1.3	Organização do Texto	5
2	FUNDAMENTAÇÃO TEÓRICA	7
2.1	Aprendizado de representações para linguagem	7
2.1.1	<i>Bag-of-words</i>	8
2.1.2	Representações distribuídas para linguagem	9
2.1.2.1	<i>Word embeddings</i>	12
2.2	Modelagem de tópicos	18
2.2.1	<i>Latent Dirichlet Allocation</i>	20
2.2.2	<i>Contextualized Topic Model</i>	23
2.2.3	<i>Embedded Topic Model</i>	25
2.3	Análise textual baseada em categorias lexicais	27
2.4	Trabalhos relacionados	29
3	UMA METODOLOGIA PARA ANÁLISE DE TÓPICOS DE DE-PRESSÃO NO REDDIT	32
3.1	Construção da base de dados	32
3.2	Limpeza e Pré-processamento de dados	35
3.3	Preparação de recursos	40
3.4	Treinamento dos modelos	42
3.5	Análise de resultados	44
4	RESULTADOS	46
4.1	Resultados no <i>corpus</i> em português	46
4.1.1	Desempenho dos modelos em português segundo a coerência	46
4.1.2	Análise qualitativa	48
4.1.3	Análise léxica	51
4.2	Resultados no <i>corpus</i> em inglês	53
4.2.1	Desempenho dos modelos em português segundo a coerência	53
4.2.2	Análise qualitativa	55
4.2.2.1	Análise qualitativa dos tópicos 0 a 7	55
4.2.2.2	Análise qualitativa dos tópicos 8 a 15	58
4.2.2.3	Análise qualitativa dos tópicos 16 a 23	62

4.2.2.4	Análise qualitativa dos tópicos 24 a 17	63
4.2.3	Análise léxica	65
4.3	Discussão geral	66
5	CONCLUSÃO	69
	REFERÊNCIAS	72
	APÊNDICES	82

1 Introdução

Ao longo das últimas décadas, as doenças mentais tornaram-se problemas de saúde pública altamente prevalentes ao redor do mundo (STEEL et al., 2014). Segundo a OMS, entre 2005 e 2015 o número estimado de pessoas com desordens depressivas no planeta aumentou em 18,4% (ORGANIZATION et al., 2017). Transtornos como ansiedade, bipolaridade e o transtorno depressivo maior (TDM) – mencionado no restante deste texto simplesmente como “depressão” ou por TDM – passaram a fazer parte das discussões sobre a sociedade moderna. Tais problemas apresentam desafios à medida que a falta de diagnóstico ou o tratamento inadequado podem acarretar severas consequências para o indivíduo em questão, como abuso de drogas, suicídio, entre outras (KANDEL; RAVEIS; DAVIES, 1991; MORTENSEN et al., 2000). Por exemplo, no caso da depressão é especialmente desafiador oferecer o tratamento adequado, pois a forma oferecida de acompanhamento pode não estar alinhada com os desejos e expectativas do indivíduo, causando assim evasão e negligência no tratamento por parte dos indivíduos depressivos (RAUE et al., 2009). Além disso, a auto-estigmatização dos indivíduos com depressão (SCHOMERUS; MATSCHINGER; ANGERMEYER, 2009), que está associada a um sentimento de inadequação na sociedade, também pode ser um fator determinante no início ou continuidade de um tratamento.

Em paralelo a esse fenômeno, as redes sociais e comunidades virtuais na *web* sofreram um *boom* vertiginoso de crescimento nos últimos anos. A Figura 1, obtida de (DATA, 2020), ilustra o crescimento das redes sociais ao longo das últimas décadas. A partir de meados da década de 2000, tais redes passaram a conectar bilhões de pessoas no mundo inteiro, proporcionando mudanças significativas na sociedade moderna. As redes passaram a influir nos mais diversos âmbitos, como no ambiente de trabalho (CAO et al., 2012), participação política (ZÚÑIGA; MOLYNEUX; ZHENG, 2014) e autoestima (VOGEL et al., 2014). Por um lado, as redes sociais ampliaram os horizontes da globalização mundial, conectando indivíduos de diversas localidades, origens e classes sociais em um ambiente único. Por outro, as redes sociais também trouxeram desafios para o estudo de saúde mental, dado que sua predominância no mundo moderno expõe questionamentos sobre seus potenciais efeitos na mente humana a longo prazo.

Diante disso, um grande número de pesquisadores atentou-se às ligações entre saúde mental e uso de redes sociais. Neste contexto, estudos associaram o uso extensivo de redes sociais a níveis maiores de ansiedade e depressão nos usuários (BARRY et al., 2017; VANNUCCI; FLANNERY; OHANNESSIAN, 2017). Além disso, pesquisas também indicaram a insatisfação com a imagem corporal, baixa autoestima e má qualidade de sono (WOODS; SCOTT, 2016; KELLY et al., 2018) como características apresentadas por indivíduos que passam muito tempo de suas vidas conectados nas redes. *Cyberbullying* e

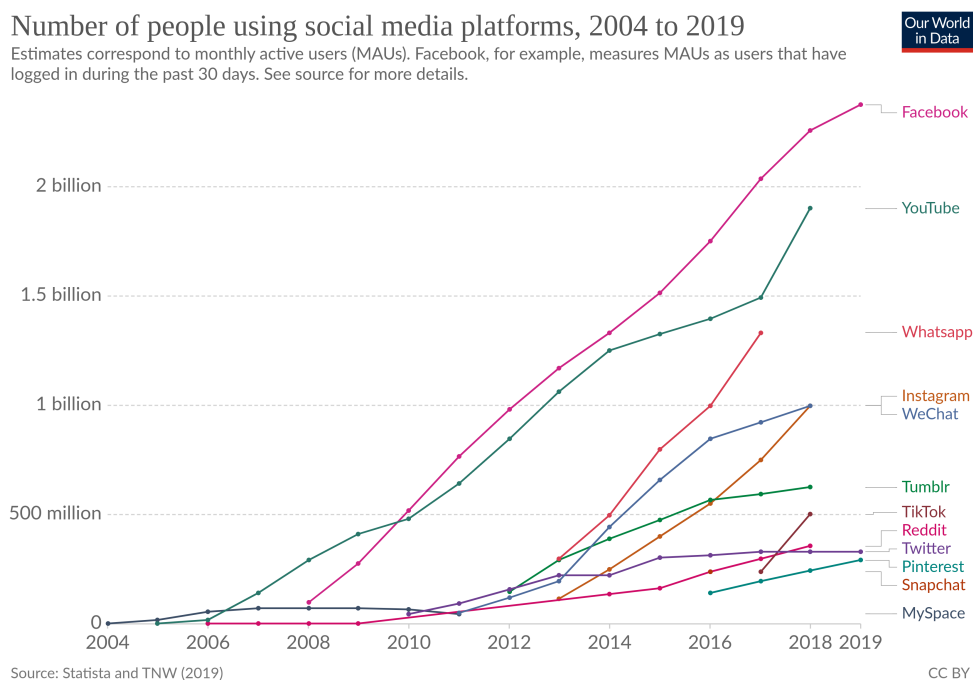


Figura 1 – Evolução do número de usuários ativos mensais em diversas redes sociais no período 2004-2019. Imagem obtida de (DATA, 2020).

campanhas de ódio também contribuem para o surgimento de sentimentos negativos em vítimas nas mídias (MISHNA et al., 2018). Sendo assim, o uso excessivo de redes sociais caracteriza-se como um fator de atenção para a comunidade médica no auxílio a indivíduos com potenciais transtornos mentais.

Em contrapartida, estudos mostram que indivíduos com variadas desordens mentais buscam as redes sociais como refúgio, seja para extravasar sentimentos ou para procurar conselhos (NASLUND et al., 2016; GKOTSIS et al., 2016). As redes oferecem um ambiente de anonimato que pode ser atraente para usuários nestas condições, incentivando a conversa aberta entre os membros das comunidades (CHOUDHURY; DE, 2014). É provável, por exemplo, que usuários depressivos se identifiquem com comunidades focadas neste tema, que incluam outros membros que passam ou já passaram pelo mesmo problema. Por vezes, indivíduos podem ter dificuldade ou medo de procurar tratamento médico para sua condição no dia-a-dia, em razão da auto-estigmatização. Sendo assim, a orientação *online* com pessoas passando por situação similar pode incentivar a procura por tratamento adequado. Dessa forma, fóruns e redes sociais podem ser uma ferramenta útil no auxílio emocional à pessoas com transtornos mentais.

Em paralelo, pesquisas atestaram que indivíduos em estado depressivo apresentam características predominantes no uso da linguagem, como o uso frequente de pronomes em primeira pessoa ou termos absolutos - como “sempre” ou “nunca” (AL-MOSAIWI; JOHNSTONE, 2018). Alguns estudos buscam correlacionar o linguajar e temas abordados

por pessoas depressivas em redes sociais, de forma a identificar temáticas relevantes para esse grupo (EICHSTAEDT et al., 2018; AL-MOSAIWI; JOHNSTONE, 2018). Tais temáticas ou tópicos poderiam indicar que uma pessoa tem tendências depressivas ou encontra-se em um estado de humor majoritariamente baixo. De forma a estudar esse processo de identificação de tópicos, técnicas de processamento de linguagem natural (MANNING; SCHUTZE, 1999) têm sido usadas e têm apresentado resultados promissores. A técnica mais empregada nesse contexto é a da modelagem de tópicos, e trabalhos anteriores já exploraram essa abordagem para auxiliar na detecção de depressão a partir de postagens na língua inglesa (RESNIK; GARRON; RESNIK, 2013; MAUPOMÉ; MEURS, 2018). Por outro lado, a técnica de análise textual automatizada de categorias lexicais também vem sendo aplicada, ainda no contexto de identificação da depressão em redes (RAMIREZ-ESPARZA et al., 2008; EICHSTAEDT et al., 2018; TADESSE et al., 2019). Dito isso, o conteúdo postado por usuários com traços depressivos em redes sociais pode guardar características linguísticas e semânticas indicativas do estado emocional no qual encontra-se seu autor.

Sendo assim, a presente monografia busca aplicar a modelagem de tópicos em um *corpus* de documentos contendo postagens de usuários em alguns sub-fóruns na rede social *Reddit*, de forma a realizar um estudo sobre o significado dos tópicos observados e sobre sua relação com temáticas indicativas de depressão. A análise de categorias lexicais presentes nos tópicos também é realizada, de forma a complementar as observações. As semelhanças e distinções características de tópicos obtidos a partir de postagens em língua portuguesa e em língua inglesa também são avaliadas. Dessa forma, busca-se encontrar os possíveis padrões que surgem na linguagem empregada pelos usuários do *Reddit* mesmo no contexto de idiomas distintos. Finalmente, a avaliação dos resultados obtidos neste estudo confirmará ou não a aplicabilidade da modelagem de tópicos para a determinação de temas relativos à depressão no contexto do *Reddit*.

1.1 Objetivos

O presente trabalho tem como um de seus objetivos a identificação dos tópicos apresentados no contexto de discussões relacionadas à depressão dentro de sub-fóruns do *Reddit*, tanto em língua inglesa quanto em português. De forma a realizar uma coleta de postagens associadas ao tema depressão, foi necessário investigar o *Reddit* em dois âmbitos: no âmbito dos sub-fóruns específicos voltados ao tema e também no âmbito dos assuntos discutidos na rede. Além de determinar quais tópicos são relevantes para discussões sobre depressão na rede, destaca-se como objetivo desta monografia a identificação das potenciais ligações entre os temas encontrados no *corpus* produzido com ambos os idiomas. Atualmente, com a alta prevalência da depressão nas diferentes regiões do planeta, é de especial importância determinar se existem tópicos relevantes em comum para a discussão

da depressão em diferentes idiomas. Além disso, pelo mesmo motivo, também torna-se relevante identificar quais são as diferenças temáticas das discussões neste contexto. Adicionalmente, o presente estudo procurou determinar, a partir dos resultados observados, a aplicabilidade dos modelos de tópicos para extração de temas latentes no contexto de uma rede social como o *Reddit*.

Entretanto, nesse contexto, a interpretação de postagens e tópicos oriundos de redes sociais sofre alguns percalços, em decorrência das características informais da linguagem empregada em tais ambientes de discussão. Abreviações, gírias, figuras de linguagem ou mesmo comunicação não-verbal – via imagens, por exemplo – são alguns dos obstáculos enfrentados ao realizar tarefas de processamento de linguagem natural (NLP) no contexto de redes e fóruns *online*. Soma-se a isso o fato do *corpus* construído com as postagens em português na rede ser substancialmente inferior ao construído para a língua inglesa, em razão da disparidade entre o número de postagens e de usuários nos sub-fóruns avaliados e entre as versões em ambos os idiomas do *Reddit*.

Conseqüentemente, as etapas de pré-processamento textual dos conjuntos de dados aqui trabalhados ganharam grande importância. Diferentes técnicas de pré-processamento foram aplicadas e avaliadas empiricamente diante dos resultados apresentados. Além disso, o mesmo conjunto de técnicas aplicado no *corpus* de um idioma não necessariamente seria o ideal a ser usado no outro, já que as diferenças entre os conjuntos tornam necessária uma avaliação criteriosa de quais configurações de pré-processamento devem ser realizadas, caso a caso.

1.2 Metodologia

Para a solução discutida na presente monografia, foi preciso realizar a coleta das postagens do *Reddit* em seu estado original em ambos os idiomas, além do pré-processamento adequado dos conjuntos de dados para permitir a sua utilização nos modelos de aprendizagem empregados. Características típicas observadas nas interações dentro do *Reddit* moldaram as escolhas realizadas em relação a que tipo de conteúdo deveria ser coletado ou não. A rede é estruturada num sistema baseado em “submissões”, que contém uma postagem inicial normalmente de tamanho mais longo seguida de uma discussão hierarquizada na forma de “respostas” de tamanho curto. Ao observar a constituição de um número significativo de submissões, percebeu-se que a maioria das respostas era formada por textos simbolizando uma reação direta ao conteúdo descrito no *post* original. Contudo, na maioria dos casos, tais respostas não buscavam aprofundar o conteúdo apresentado na postagem original. Diante dessas características e da estrutura centralizada na postagem original com a qual a rede hierarquiza as discussões ali existentes, decidiu-se que as postagens originais seriam as mais significativas para determinar os temas discutidos

dentro do *Reddit*. Portanto, a construção dos *corpora* aqui discutidos concentrou-se no conteúdo das postagens originais. Em relação ao procedimento de pré-processamento, de forma a determinar a melhor abordagem para cada um dos *corpora*, foi preciso realizar experimentos aplicando diferentes métodos de forma empírica.

Os resultados foram avaliados no contexto dos modelos treinados a partir de cada *corpus* pré-processado. Como diferentes modelos de aprendizagem de tópicos foram explorados, foi necessário avaliar os resultados para cada um deles. Para tal, foram necessárias etapas de preparação adicionais específicas para cada tipo de modelo a ser treinado, já que cada um deles possui arquitetura, interface e saídas distintas. Na presente monografia, foram explorados os seguintes modelos: *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003), um modelo probabilístico e o mais popular para modelagem de tópicos; *Contextualized Topic Model* (CTM) (BIANCHI et al., 2021), que emprega representações pré-treinadas de linguagem como o SBERT para aprender os tópicos; e *Embedded Topic Model* (ETM) (DIENG; RUIZ; BLEI, 2020), um modelo baseado em redes neurais e em representações de palavras usando *embeddings word2vec*.

Os resultados obtidos foram avaliados quantitativamente a partir do emprego da métrica de coerência, usada para determinar em modelos não-supervisionados aqueles com maior qualidade de tópicos. Além disso, apesar das distinções de arquitetura e de modelo de dados para entrada entre os diferentes modelos, o cálculo da métrica de coerência foi unificado entre os modelos, de forma a obter maior uniformidade entre os valores apresentados, proporcionando uma análise mais precisa da qualidade dos diferentes tópicos.

A seguir, uma análise qualitativa foi realizada, a partir da leitura e rotulação manual dos tópicos produzidos por cada modelo. Ainda, uma análise das categorias lexicais representadas pelas palavras associadas aos tópicos gerados também foi feita. Dessa forma, pôde-se determinar a qualidade dos tópicos extraídos pelos modelos, confirmando ou não a aplicabilidade da técnica de modelagem de tópicos dentro do contexto do *Reddit*. Por fim, uma discussão sobre os tópicos aprendidos pelos modelos nos *corpora* dos dois idiomas aqui estudados foi realizada. As semelhanças e diferenças entre os tópicos em português e em inglês foram notadas, de forma a caracterizar se há ou não uma possível ligação nos temas abordados por falantes de ambas as línguas no contexto da depressão.

1.3 Organização do Texto

O restante desta monografia está organizado como descrito a seguir. No capítulo 2, descreve-se a fundamentação teórica da presente monografia, elucidando o contexto no qual o trabalho encontra-se, as áreas de estudo envolvidas e discutindo demais trabalhos na literatura que abordaram temas similares ou relacionados ao tratado pelo presente texto.

No capítulo 3 descreve-se a metodologia aplicada para solução do problema proposto, abordando desde a produção dos *corpora* estudados até a estruturação do treinamento de cada um dos modelos utilizados, finalizando com a definição da etapa de análise dos resultados. Em seguida, o capítulo 4 aborda os resultados obtidos mediante o treinamento discutido. Finalmente, o capítulo 5 discorre acerca das observações finais proporcionadas pelo estudo, reforçando os achados obtidos e propondo caminhos de estudo futuros.

2 Fundamentação Teórica

Para o presente capítulo, objetiva-se introduzir o contexto no qual esta monografia se insere no âmbito da inteligência artificial. Especificamente, o presente trabalho emprega técnicas associadas ao processamento de linguagem natural (NLP) e ao aprendizado de máquina para realização dos objetivos discutidos. As técnicas, áreas de estudo e conceitos relacionados diretamente ao trabalho são descritas nas seções a seguir. Primeiramente, aborda-se as técnicas para representação de linguagem textual empregadas para tarefas de NLP, como as deste estudo. Em seguida, é discutida a área de modelagem de tópicos, responsável pela extração dos temas relevantes de um dado conjunto textual. Prossegue-se com uma discussão sobre análise textual baseada em categorias lexicais, que também foi uma técnica explorada neste trabalho. Por fim, esta seção discute trabalhos de literatura relacionada que exploram a extração de tópicos e a modelagem de linguagem num contexto de identificação das características linguísticas associadas às discussões *online* sobre problemas mentais.

2.1 Aprendizado de representações para linguagem

A linguagem humana, em suas mais variadas formas, é dotada das mais diversas complexidades e características, e sua imensa variedade frequentemente tem relação com o contexto social, cultural e geográfico em que encontra-se (ASSUNÇÃO, 2010). Ao longo do tempo, a linguagem tem se adaptado aos costumes e acontecimentos que permeiam a sociedade humana. Ainda, elementos como regionalismos, figuras de linguagem, metáforas, entre outros, enriquecem a expressividade linguística humana. Contudo, tais características voláteis, subjetivas e complexas da linguagem podem representar uma dificuldade quando deseja-se que computadores e seus algoritmos compreendam um conjunto textual de documentos – comumente chamado de *corpus* – e extraiam informações precisas do mesmo. Isto deve-se ao fato de que computadores não conhecem o contexto nem possuem toda a base de informação necessária para a compreensão de uma frase ou texto composto por ambiguidades e semântica complexa (FALLER, 2020).

Neste contexto, insere-se a área de processamento de linguagem natural (NLP), uma subárea da inteligência artificial. A área de NLP tem por objetivo estudar a representação e análise de conjuntos textuais heterogêneos, empregando para tal ferramentas computacionais que aproximam o processamento e compreensão da linguagem realizado por humanos (LIDDY, 2001). Comumente, técnicas de NLP são utilizadas como ferramenta para solução de problemas dos mais diversos tipos, como extração e recuperação de informações, tradução de máquina, entre outros (LEE, 2020).

No presente estudo, objetiva-se realizar a extração de tópicos a partir de um *corpus* textual oriundo de postagens de rede social. A tarefa de modelagem de tópicos também é uma das atividades estudadas dentro de NLP. Para sua realização, deve-se determinar as formas de representação da linguagem a serem utilizadas como entrada para os diferentes tipos de modelos de tópicos. Cada forma de representação tem características próprias e acarretará na percepção de diferentes nuances textuais por parte de cada modelo. Deve-se notar que a troca de um método de representação por outro pode provocar diferenças significativas nos resultados obtidos, dado que diferentes estruturas de dados adaptam-se melhor a determinados problemas do que a outros (BARUA, 2020). As subseções a seguir tratam sobre populares formas de representação para tarefas de linguagem a serem resolvidas com aprendizado de máquina (MITCHELL, 1997) e empregadas no presente trabalho. A primeira subseção aborda a forma de representação em *bag-of-words* (BOW) (HARRIS, 1954), que é uma representação local da linguagem. Em seguida, discutem-se as características das representações distribuídas, que também foram exploradas neste estudo e que possuem vantagens em diferentes aspectos em relação às representações locais como o BOW. Na mesma subseção, há um aprofundamento em duas formas de representação distribuída empregadas no trabalho: *word2vec* (MIKOLOV et al., 2013b) e BERT (DEVLIN et al., 2019).

2.1.1 *Bag-of-words*

O modelo *bag-of-words* (BOW) - do inglês “saco de palavras” - busca representar de forma conceitualmente simples um dado conjunto de termos. Uma das primeiras referências ao “saco de palavras” num contexto de linguística foi realizada por Harris em 1954 (HARRIS, 1954). Para um dado texto, a representação em BOW é realizada por meio da construção de um conjunto sem duplicatas dos termos que ocorrem no referido documento. A seguir, para cada termo do conjunto criado é associado um valor para o termo no texto, comumente a sua contagem ou frequência no documento. O conjunto de pares termo-contagem ou termo-frequência assim criado é o chamado *bag-of-words* do documento, e seu nome tem origem na falta de ordenação dos termos representados ao usar tal modelo. O modelo BOW é um exemplo da chamada **representação local** ou baseada em símbolos para palavras, pois cada valor na representação gerada mapeia uma determinada entrada (LIU; LIN; SUN, 2020). Por fim, o BOW produzido pode ser então utilizado para realizar a representação vetorial dos termos e documentos a serem avaliados. Na Figura 2, pode-se observar a codificação BOW de um dado documento. Note que no exemplo a ordem do vocabulário foi definida como a ordem em que os termos únicos aparecem no texto. Esta ordem define também a representação vetorial produzida, onde cada posição representa um dos termos do vocabulário e seu valor representa a contagem de vezes que tal termo aparece no documento.

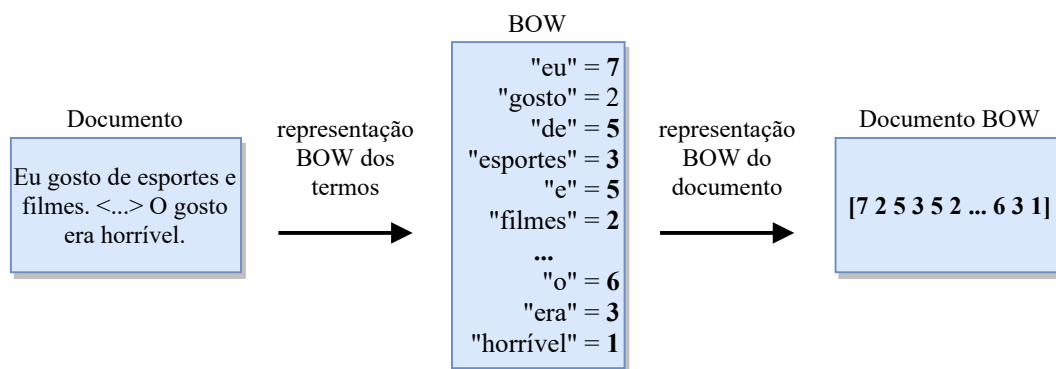


Figura 2 – Codificação BOW dos termos de um documento e do próprio documento.

Um dos problemas associados ao BOW é o fato de que ele mantém apenas a informação de contagem de termos, descartando informações sobre ordenação de palavras. O modelo BOW sabe apenas se termos ocorreram no texto, mas não sabe quando nem como ocorreram (BROWNLE, 2017a). Em razão disso, tal modelo não absorve informações de semântica e contexto de palavras. Por exemplo, na Figura 2, a palavra “gosto” é empregada em dois contextos distintos com significado adverso, mas como o BOW armazena apenas a contagem de termos, ambas as ocorrências são consideradas equivalentes. Além disso, nota-se que para representar um determinado documento por meio de um vetor, a representação em BOW do mesmo deve conter uma entrada para cada um dos termos existentes no *corpus*. Desta forma, o vetor v que representa um documento d pertencerá ao espaço $\mathbb{R}^{|V|}$, onde $|V|$ é o tamanho do vocabulário do *corpus*. Como nem todos os termos aparecem em todos os documentos, no caso de vocabulários cada vez mais amplos, existirão documentos em BOW cujas representações vetoriais estarão repletas de zeros para as entradas de termos que não aparecem nos documentos. Desta forma, a representação BOW de tal *corpus* de documentos será composta por uma matriz esparsa, onde cada linha representa um determinado documento. Neste contexto, a alta esparsidade de matrizes representa mais um problema do modelo BOW. O problema ocorre pois múltiplas entradas zeradas numa matriz esparsa representam a inexistência de dados suficientes em muitas das amostras ou documentos observados. Em razão desta característica, a capacidade do modelo em determinar padrões e descobrir distribuições no *corpus* será prejudicada (ALLISON; GUTHRIE; GUTHRIE, 2006). De forma a prevenir problemas como estes, surgiu como alternativa a abordagem de aprendizado de **representações distribuídas** para linguagem.

2.1.2 Representações distribuídas para linguagem

As representações distribuídas representam de forma compacta um *corpus*, usualmente reduzindo a dimensionalidade do conjunto trabalhado de forma a combater o chamado “*curse of dimensionality*”: o fato de que a sequência de termos encontrados por

um modelo em sua fase de testes provavelmente será totalmente distinta das sequências de termos encontradas pelo mesmo modelo na etapa de treinamento (BENGIO et al., 2003). O problema acontece tipicamente em representações locais como o BOW, onde o conjunto de termos de um *corpus* é modelado segundo variáveis discretas. Em um espaço de variáveis discretas, qualquer alteração nos valores das variáveis pode alterar drasticamente as distribuições sendo estimadas. Isto inclui as distinções na sequência de termos encontrados por um modelo em suas fases de treinamento e testes. Desta forma, matrizes esparsas como as produzidas usando BOW estão sujeitas a este problema, dado o grande potencial de variabilidade em suas entradas. Enquanto isso, representações distribuídas mapeiam os termos em um espaço contínuo de variáveis, onde perturbações em seus valores terão impactos menores e mais previsíveis sob as distribuições mapeadas. Desta forma, esse tipo de representação possibilita como benefício em relação ao BOW uma maior robustez ao lidar com o problema da esparsidade de dados (LIU; LIN; SUN, 2020). A principal hipótese por trás das representações distribuídas é a de que “*uma palavra é caracterizada pela companhia que tem*” (FIRTH, 1957). Sendo assim, este tipo de representação busca modelar de forma semelhante termos que possuam semântica parecida. Conseqüentemente, essas representações também conseguem lidar melhor com o problema da distinção semântica entre termos baseada no contexto de emprego das palavras.

Enquanto no BOW há uma correspondência um-para-um entre a representação gerada e o dado original, nas representações distribuídas cada dado original é representado por um padrão de ativação distribuído entre diversos elementos de computação (LIU; LIN; SUN, 2020). Além disso, cada elemento de computação também será responsável por auxiliar na representação dos demais dados. Em decorrência dessas características, este modelo recebe seu nome. O mapeamento dos dados originais em representações distribuídas é comumente aprendido por meio de algoritmos envolvendo redes neurais artificiais e aprendizado profundo. Assim, as representações facilitam a transferência de conhecimento entre diferentes tarefas. Pode-se treinar uma representação distribuída em determinada tarefa de processamento de linguagem, e então utilizar as representações aprendidas para solucionar outro problema (LIU; LIN; SUN, 2020). A utilização de representações distribuídas para NLP possibilitou o surgimento dos chamados *word embeddings*, que são representações de palavras na forma de vetores com valores reais, e que viabilizaram uma revolução na área de NLP (BARUA, 2020). A seguir, serão descritos os conceitos básicos associados às redes neurais artificiais (ANNs), utilizadas no aprendizado de *word embeddings*. Posteriormente, há uma discussão sobre o que são os *word embeddings* e sobre as versões do mesmo empregadas no presente estudo.

Redes neurais artificiais

As redes neurais artificiais, do inglês “*artificial neural networks*” (ANNs) são um conjunto de algoritmos em inteligência artificial que buscam simular aproximadamente o

comportamento dos processos cerebrais humanos para realização de aprendizagem. Em princípio, as ANNs simulam as conexões neuronais do cérebro, modelando a troca de informações entre unidades de processamento chamadas de neurônios artificiais, responsáveis pela produção de um resultado – normalmente numérico ou categórico – a partir de um conjunto de entradas. As origens dos dados de entrada numa rede neural são variáveis – dados textuais, numéricos, imagens, sons e vídeos podem ser empregados, desde que devidamente codificados de forma numérica.

As ANNs podem ter diversas arquiteturas diferentes, como camada única ou múltiplas camadas de neurônios. Comumente, a arquitetura com múltiplas camadas é utilizada, uma vez que desta forma uma ANN é um aproximador universal de funções (WHITE, 1992). Neste caso, as ANNs são compostas de ao menos três camadas: a camada de entrada, que recebe os dados de entrada; a camada escondida, que processa os dados de entrada por meio de sua combinação, definindo uma saída; e a camada de saída, que define os resultados da rede neural a partir das entradas utilizadas. Pode-se ter múltiplas camadas escondidas numa ANN, onde cada camada pode ser responsável pelo aprendizado de um conhecimento específico necessário para a solução do problema em questão. Cada uma das camadas é formada por unidades chamadas neurônios, que a partir de determinada combinação dos dados de entrada produzirão suas respectivas saídas. Além disso, cada conexão numa ANN terá um peso associado, determinado de acordo com sua importância e aprendido conforme o treinamento da rede. A saída de um neurônio é resultante da aplicação de uma função de ativação sobre a combinação linear dos dados de entrada e dos pesos. O objetivo ao utilizar ANNs é o de aproximar, por meio da rede, o mapeamento entre um determinado conjunto de entradas em uma determinada saída ou conjunto de saídas. Pode-se objetivar mapear imagens em categorias de objetos (PATHAK; PANDEY; RAUTARAY, 2018), mapear obras textuais em gêneros literários (WORSHAM; KALITA, 2018), assim como mapear palavras em representações vetoriais – o que é realizado ao treinar-se *word embeddings*, por exemplo (MIKOLOV et al., 2013b; DEVLIN et al., 2019).

A Figura 3 mostra uma ANN completamente conectada com apenas uma camada escondida para uma tarefa de classificação de imagens. Uma imagem é dada como entrada para a rede e a classificação é realizada, determinando a constituição da figura a partir de um conjunto de probabilidades. A fotografia de cavalos foi redimensionada a partir da original por (MARCHAL, 2005).

As ANNs, em suas diferentes variações, foram aplicadas com sucesso em diversos cenários, como classificação de texto (LAI et al., 2015), geração de texto (BROWN et al., 2020), geração de imagens (RAMESH et al., 2021), jogos competitivos (SILVER et al., 2017; VINYALS et al., 2019), reconhecimento de escrita (BALDOMINOS; SAEZ; ISASI, 2018), filtragem de *spams* (CLARK; KOPRINSKA; POON, 2003), entre outros. Diante disso, pesquisadores de NLP também decidiram empregar as ANNs para o reconhecimento

de padrões que ajudassem a resolver tarefas como o aprendizado de linguagem. Como exemplo, os modelos de linguagem *word2vec* e BERT empregam ANNs em sua constituição arquitetural, de forma a realizar o aprendizado de representações de palavras, e trouxeram avanços significativos à área.

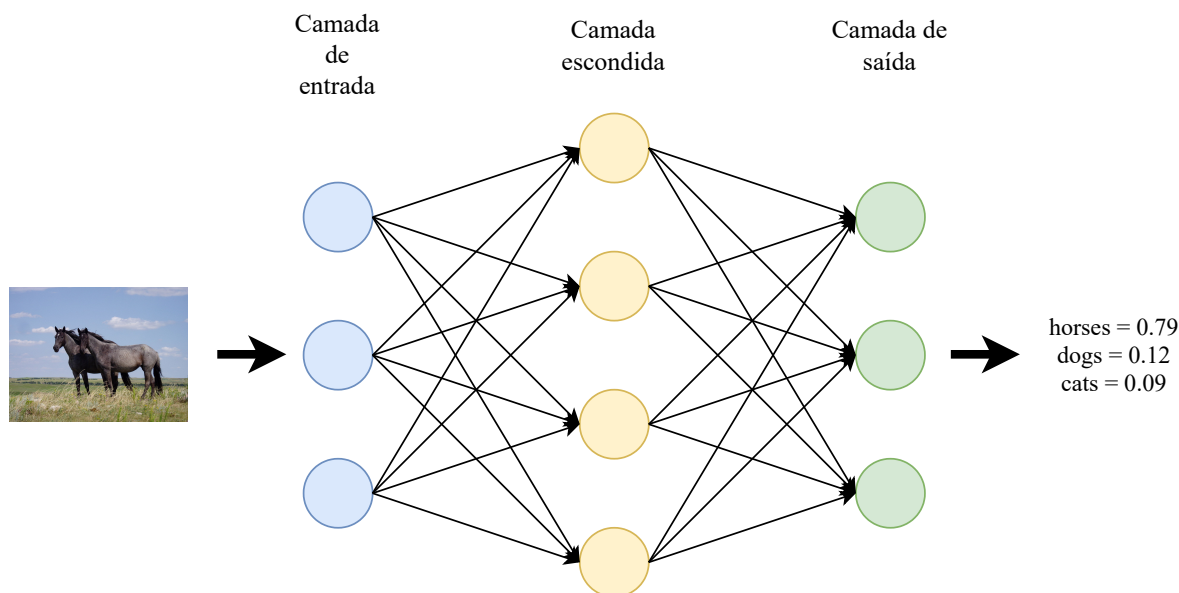


Figura 3 – Arquitetura de uma rede neural artificial para classificação de imagens. Utiliza fotografia obtida de (MARCHAL, 2005).

2.1.2.1 Word embeddings

Word embeddings é o nome dado a uma classe de técnicas de representação textual distribuída baseadas no aprendizado de representações via ANNs ou modelos estatísticos. Os métodos que aprendem *word embeddings* induzem representações distribuídas densas dos termos presentes em um *corpus* (BROWNLE, 2017b). Os termos são representados na forma de vetores densos de tamanho predefinido compostos por números reais. Sendo assim, a representação vetorial deixa de crescer à medida que o vocabulário é incrementado como na representação *BOW* e passa a ser um hiperparâmetro de aprendizagem do modelo. Os *word embeddings* são aprendidos por meio de um algoritmo de aprendizado específico, como o *Word2Vec* (MIKOLOV et al., 2013a), *GloVe* (PENNINGTON; SOCHER; MANNING, 2014), BERT (DEVLIN et al., 2018), entre outros. Os algoritmos aprendem os vetores de cada termo a partir do contexto de uso observado no *corpus*.

Diante do descrito, *word embeddings* viabilizam a representação mais rica de um *corpus*, capturando semântica e contexto de forma mais apropriada que representações *bag-of-words* (TOSHEVSKA; STOJANOVSKA; KALAJDJIESKI, 2020), além de proporcionarem elementos característicos das representações distribuídas como redução de dimensionalidade e possibilidade de transferência de conhecimento (LIU; LIN; SUN, 2020).

Existem diversas técnicas de aprendizagem para *word embeddings*, apropriadas para diferentes contextos de uso. No presente trabalho, o foco será nas abordagens empregadas pelo *Word2Vec* e pelo BERT. A principal diferença entre ambos os modelos está no fato de que o *word2vec* aprende apenas uma representação vetorial densa para cada palavra do *corpus* de treinamento, enquanto um mesmo termo poderá possuir múltiplas representações distintas no BERT, a depender de seu contexto de uso.

O algoritmo *Word2Vec* foi criado por Mikolov et al (MIKOLOV et al., 2013a), sendo projetado para tornar mais eficiente o treinamento de *embeddings* a partir de ANNs. Ao longo dos últimos anos, o *word2Vec* tornou-se uma das formas mais populares para treinamento de *word embeddings* a serem utilizados em diferentes tarefas de processamento de linguagem. O modelo *word2Vec* tem uma arquitetura de aprendizado auto-supervisionado (BURKOV, 2019).

O modelo *word2Vec* é capaz de absorver regularidades semânticas e sintáticas na linguagem (MIKOLOV; YIH; ZWEIG, 2013), como a relação intrínseca entre os termos “homem” e “mulher”. Além disso, mostrou-se que a aritmética baseada nos vetores *word2Vec* produz resultados semanticamente coerentes com o significado dos vetores envolvidos. Por exemplo, percebeu-se que, dados os vetores que representam os termos “rei”, “homem”, “rainha” e “mulher”, ao realizar-se a subtração “rei” - “homem” + “mulher” o resultado obtido é de um vetor espacialmente próximo ao vetor “rainha”. A Figura 4 mostra a representação vetorial em *word2vec* dos termos de um documento. Note que termos que ocorrem em contextos similares têm representações vetoriais semelhantes, como pode-se perceber nas representações dos termos da primeira frase do documento. Contudo, existe apenas uma única representação para cada termo existente no *corpus*.

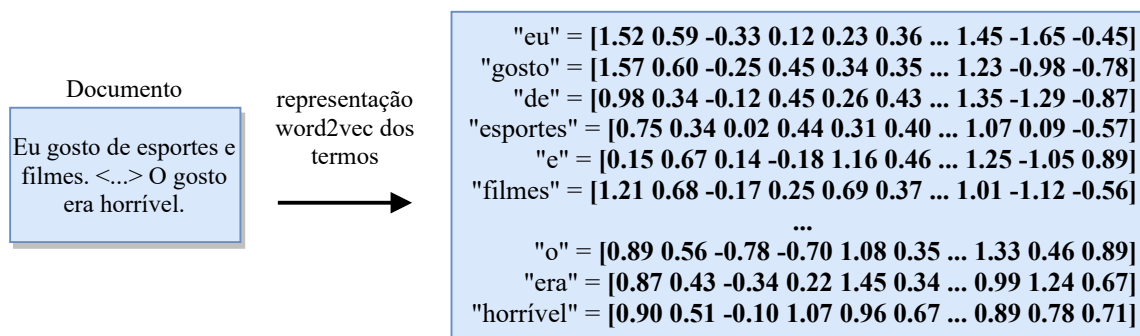


Figura 4 – Codificação *word2vec* dos termos de um documento.

Os modelos *word2Vec* subdividem-se em duas abordagens que diferem na forma de enxergar o problema a ser solucionado: *continuous bag-of-words* (CBOW), que busca aprender a representação vetorial de um termo a partir das palavras em seu contexto; e *skipgram*, que busca aprender a representação vetorial de um conjunto de termos representando um contexto a partir de uma palavra-alvo fornecida. Em ambas as abordagens, o contexto de uma palavra é delimitado pelo conceito de janela de palavras. Uma janela de

palavra inclui as palavras precedentes e sucedentes à palavra-alvo atual. O tamanho de tal janela, ou seja, quantos elementos ela incluirá, é um hiperparâmetro do modelo. Conforme o modelo percorre o conjunto de termos a serem aprendidos, a janela de palavras também desloca-se.

A Figura 5 exibe as arquiteturas CBOW e *skipgram*. Note que a arquitetura CBOW recebe um conjunto de representações de palavras – um contexto – e determina a representação *word2vec* de uma palavra-alvo. Enquanto isso, a arquitetura *skipgram* recebe a representação de uma palavra e determina a representação *word2vec* dos termos presentes no contexto desta palavra-alvo. Perceba que as arquiteturas têm estruturação inversa entre si.

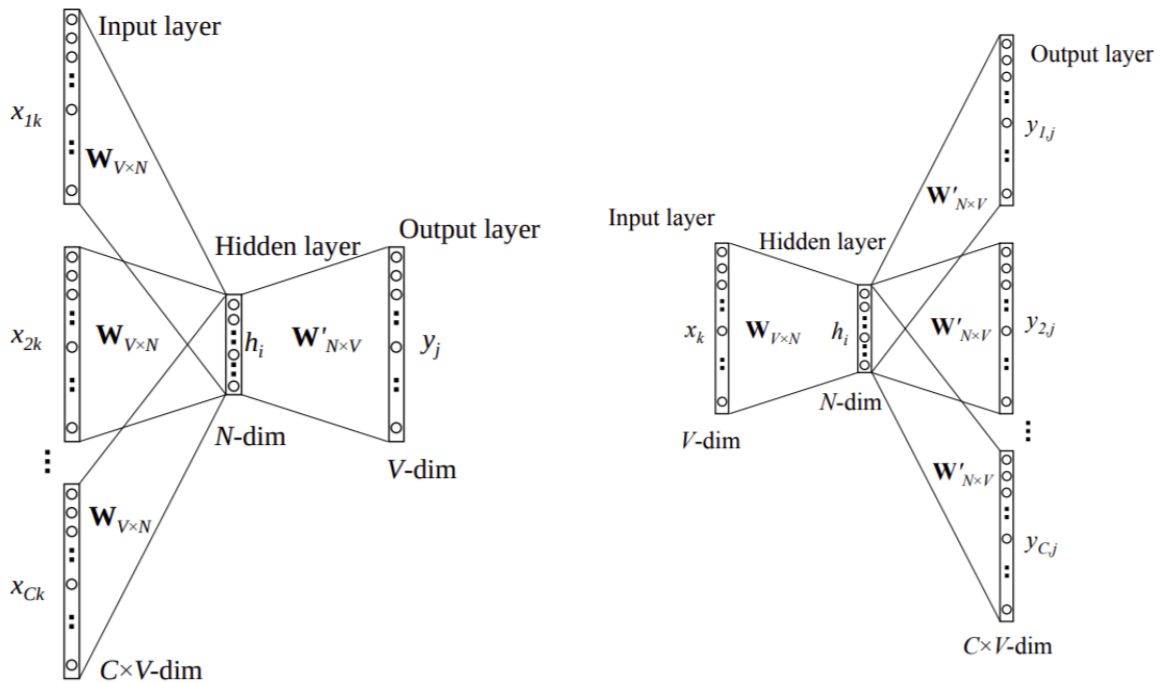


Figura 5 – Arquiteturas dos modelos CBOW e *continuous skipgram*, obtidas de (RONG, 2014).

Em razão da existência de apenas uma representação vetorial para cada termo de um *corpus* no *word2vec*, o modelo é sujeito a ter dificuldade para distinguir termos polissêmicos, por exemplo. Considere as seguintes frases: “Eu gosto de assistir séries” e “A função do acompanhante era assistir o paciente”. O *word2vec* falhará em distinguir semanticamente os dois significados distintos da palavra “assistir” nas frases anteriores, exatamente em razão de que aprenderá apenas uma representação do referido termo. Para solucionar problemas como este, surgiram métodos que aprendem *word embeddings* contextualizados (PETERS et al., 2018; DEVLIN et al., 2019). Em tais abordagens, um mesmo termo poderá possuir diferentes representações a depender de seu contexto de uso. No contexto de *word embeddings* contextualizados, o foco desta monografia será no

algoritmo nomeado de BERT, uma vez que o mesmo será empregado no presente trabalho.

O algoritmo BERT, do inglês “*Bidirectional Encoder Representations from Transformers*” é um modelo de linguagem pré-treinado para tarefas de NLP baseado na arquitetura de *Transformers* (VASWANI et al., 2017). Idealizado em (DEVLIN et al., 2018), o BERT baseia-se em uma arquitetura bidirecional, que leva em consideração os termos que precedem e sucedem uma palavra a ser aprendida, mediante o emprego de um delimitador de contexto de palavras. Diferentemente da representação aprendida pelo *word2vec*, onde uma palavra terá sempre a mesma representação vetorial mesmo que usada em contextos diferentes, no BERT o vetor de um termo terá constituição distinta conforme o contexto. Em consequência da forma de aprendizado do BERT, o modelo consegue aprender palavras considerando o contexto ao redor das mesmas. No exemplo citado anteriormente, onde a palavra “assistir” era usada nas frases “Eu gosto de assistir séries” e “A função do acompanhante era assistir o paciente”, o BERT conseguiria distinguir os diferentes usos da palavra por meio de representações vetoriais adequadas.

O aprendizado no BERT é realizado com o uso do *Transformer* (VASWANI et al., 2017), uma arquitetura de rede neural baseada no mecanismo de atenção que aprende relações de contexto entre elementos de um dado conjunto de entrada, considerando ou não dependências temporais entre os elementos. No caso do BERT, o *Transformer* é usado para o aprendizado das relações de contexto entre as palavras presentes em um texto e considera a dependência temporal entre os *tokens* a serem aprendidos. Entretanto, enquanto na arquitetura do *Transformer* original existem um *encoder* e um *decoder*, no BERT é preciso apenas do *encoder*, pois o objetivo aqui limita-se a obter uma representação de linguagem. Ademais, o BERT foi idealizado com base em conceitos e modelos apresentados em trabalhos sobre transferência de aprendizado – do inglês *transfer learning* (TL) – como (PETERS et al., 2018; HOWARD; RUDER, 2018; RADFORD et al., 2018), além do já referido *Transformer*. Desta forma, a ideia por trás do modelo é realizar o aprendizado das representações vetoriais a partir de um *corpus* amplo, posteriormente utilizando o conhecimento adquirido em outra tarefa de NLP com propósito específico.

O *input* do modelo BERT é uma sequência de termos representados por vetores. Tal sequência representa uma sentença. As sentenças no BERT têm uma limitação de tamanho de 512 *tokens* de entrada (DEVLIN et al., 2018) em razão do custo computacional de treinamento do modelo. Caso as sentenças possuam uma quantidade superior de termos, apenas os primeiros 512 serão mantidos. Além dos termos da sequência, são introduzidos também na entrada *tokens* especiais utilizados nas tarefas de aprendizado do modelo, a depender da tarefa realizada em dado momento. Os *tokens* são os seguintes:

- *[MASK]* - *token* máscara, substitui um dos termos da sentença ou sequência original. Usado na tarefa de *masked language model* (MLM);

- $[CLS]$ - *token* indicativo do início de uma sentença ou sequência de termos. Usado na tarefa de *next sentence prediction* (NSP);
- $[SEP]$ - *token* indicativo do fim de uma sentença ou sequência de termos. Usado na tarefa de *next sentence prediction* (NSP).

Os vetores serão processados por uma rede neural que produzirá como saída um conjunto de vetores, onde cada vetor representará um dos termos de entrada, na mesma ordem. A Figura 6 (DEVLIN et al., 2018) é uma ilustração em alto-nível da arquitetura do BERT, e evidencia a característica bidirecional do modelo por meio das camadas escondidas completamente conectadas da figura. Além disso, a Figura 7 exemplifica a transformação de um conjunto de termos de um documento em sua respectiva representação BERT. Note que termos que ocorrem em contextos semelhantes têm representação vetorial similar. Recapitulando o exemplo da palavra “assistir” citado anteriormente, no BERT caso tal termo apareça duas vezes em contextos distintos no *corpus*, então o mesmo possuirá duas representações vetoriais distintas, cada uma representando um contexto de ocorrência. Dessa forma, um termo pode ter várias representações distintas no BERT, dependendo de sua semântica e contexto de uso.

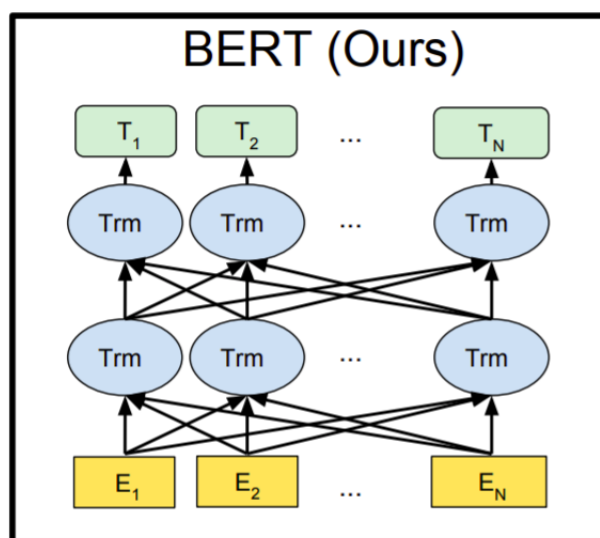


Figura 6 – Arquitetura do BERT em alto-nível, obtida de (DEVLIN et al., 2018).

O artigo original sobre o BERT introduziu duas arquiteturas: o BERT *base*, com dimensões menores e o BERT *large*, de maior dimensão e aplicado com sucesso a diversas tarefas de NLP no trabalho supracitado (SETH, 2019). De forma geral, o BERT é pré-treinado em duas tarefas de aprendizado distintas para facilitar a aprendizagem de contexto. Essas duas tarefas são modelagem de linguagem mascarada – *masked language model* (MLM) – e predição da próxima sentença – *next sentence prediction* (NSP). Ambas as tarefas são realizadas simultaneamente durante o treinamento do BERT.

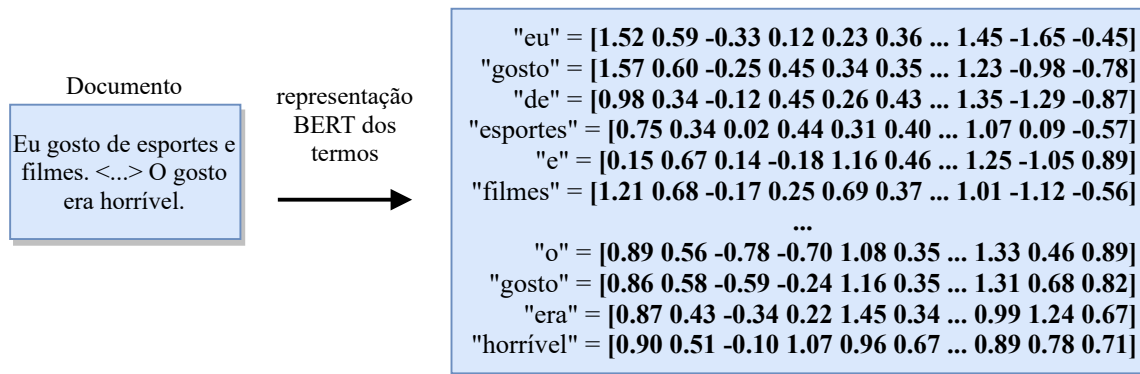


Figura 7 – Codificação BERT dos termos de um documento.

- **Masked Language Model:** Na tarefa de aprendizado MLM, o objetivo é determinar qual palavra um *token* de máscara inserido na entrada substituiu na sequência de termos original. Nesta tarefa, antes da entrada de dados, 15% dos termos de uma sequência são substituídos por um *token* [MASK]. Esta sequência, incluindo os *tokens* “mascarados”, é então utilizada como entrada do BERT. O modelo, por sua vez, irá determinar quais foram os termos substituídos da sentença original. Desta forma, o que o modelo BERT fará nesta etapa será aprender a representação dos termos substituídos. A substituição de 15% dos *tokens* de entrada não é a única manipulação da entrada disponível ao BERT, pode-se por exemplo substituir um determinado termo da sentença por outro em diferente posição e, de maneira similar, inferir qual termo originalmente ocupava a posição modificada (ALAMMAR, 2018). A saída do modelo, nesta tarefa, será usada para alimentar um classificador, cujo resultado produzido via função *softmax* determinará a probabilidade de um termo do *corpus* ser o termo original substituído pelo *token* “máscara”, para cada uma das posições onde houve tal troca.
- **Next Sentence Prediction:** Na tarefa NSP, o objetivo é determinar se uma sequência de termos *B* sucede diretamente uma sequência *A*, dadas duas sequências de termos *A* e *B* quaisquer utilizadas como entrada (HOREV, 2018). Nesta tarefa, o treinamento é realizado da seguinte forma: 50% das vezes a sequência *B* será sucessora da sequência *A* dada, enquanto 50% das vezes isto não será verdade, i.e. a sequência *B* será uma sentença qualquer obtida de alguma outra parte do *corpus*. A entrada nesta tarefa deve indicar o início e o fim de cada sequência, feito respectivamente por meio da inserção dos *tokens* [CLS] e [SEP] em cada sentença. Ainda, deve-se indicar o pertencimento de cada *token* a cada sequência, realizado pela adição de um *embedding* de sentença em cada termo. Por fim, um *embedding* posicional é adicionado a cada um dos termos, representando a posição do termo na sentença respectiva. Portanto, a entrada em tal formato é aplicada no modelo, que irá por meio de um classificador determinar se a sequência *B* é sucessora de *A*

ou não. Novamente, o resultado do classificador é produzido com o uso da função *softmax*, produzindo uma probabilidade para ambas as possibilidades.

O modelo BERT inspirou múltiplas variantes com diferentes objetivos e otimizações (DEVOPEDIA, 2019). Uma das variantes do BERT é o SBERT, proposto em (REIMERS; GUREVYCH, 2019). O BERT é computacionalmente muito custoso para realização de tarefas de similaridade semântica entre pares de sentenças, em razão do fato de que a arquitetura precisa processar duas sentenças separadamente de forma a obter as informações necessárias para executar comparações entre as mesmas (KARLSSON, 2020). A principal motivação do SBERT é a realização mais veloz de comparações entre sentenças, já que o modelo emprega estruturas de redes siamesas e triplas para realização de processamento em conjunto de múltiplas sentenças. Dessa forma, o SBERT viabiliza com maior eficiência a representação de sentenças usando *embeddings*. No presente trabalho, o SBERT foi especificamente utilizado como uma das formas de modelagem textual usando *embeddings*.

2.2 Modelagem de tópicos

Na atual era digital, enormes quantidades de dados brutos são geradas diariamente nas mais variadas plataformas da *web*, como redes sociais, artigos em *blogs* e jornais eletrônicos, vídeo e música via *streaming*, entre outros. O crescimento da quantidade de informações disponibilizadas em formato digital afetou diferentes setores, desde governos até as artes, naturalmente incluindo a ciência (PRESS, 2013). Neste contexto, tal proliferação de dados impõe desafios quanto à capacidade de categorização e extração de informações valiosas desses dados. Para gerenciar a explosão de dados, faz-se necessário o emprego de técnicas que sejam capazes de organizar, indexar, pesquisar e navegar largas coleções de documentos eletrônicos (ALGHAMDI; ALFALQI, 2015).

Diante disto, a organização das informações em categorias tornou-se de enorme importância para pesquisadores dos mais diferentes ramos (KECHAGIA, 2015). As informações de origem textual estão inclusas neste conjunto de interesse, abarcando uma grande diversidade de conteúdo, oriundo de artigos acadêmicos e obras literárias, passando por textos jornalísticos e postagens de redes sociais. Em razão da enorme proporção de dados trafegados na internet e sua ampla heterogeneidade, é impossível realizar anotações ou categorizações de forma manual em toda a massa de informações. Nesse contexto, técnicas de processamento de linguagem natural, como a modelagem de tópicos, mostram-se uma opção essencial para extrair tendências textuais de forma não-supervisionada. A categorização textual no ambiente digital auxilia a compreensão do *corpora* textual, pois viabiliza a percepção de tendências e padrões em dados de grande escala (CURISKIS et al., 2020).

A modelagem de tópicos consiste de, a partir de um *corpus*, determinar os tópicos ou assuntos que permeiam tal conjunto, identificando os seus temas mais proeminentes. Um

tópico é definido por uma coleção de palavras que têm significado semelhante ou contexto de uso parecido. No cenário ideal, em um tópico sobre “meio-ambiente” por exemplo, termos como “Amazônia”, “queimadas”, “poluição” ou “desflorestamento” apareceriam associados ao tema. Deve-se notar que na modelagem de tópicos, apenas os conjuntos de palavras que representam os temas e alguma função de pertencimento são descobertos – por isso, eles são chamados de “tópicos latentes” (MOHR; BOGDANOV, 2013). Desta forma, é de grande importância que os tópicos tenham semântica clara, para que futuras rotulações do tópico sejam realizadas adequadamente.

Uma ampla gama de aplicações foi realizada por meio de modelagem de tópicos. Quanto ao tipo de *corpus* analisado, a modelagem de tópicos foi aplicada em conjuntos textuais sobre notícias jornalísticas, lendas folclóricas (TANGHERLINI; LEONARD, 2013), documentos governamentais de diferentes eras históricas (MILLER, 2013; MOHR et al., 2013), dissertações acadêmicas (CHUANG et al., 2012), postagens em redes sociais (OSTROWSKI, 2015) e obras literárias (SCHÖCH, 2017). Além disso, a modelagem de tópicos também foi adaptada para áreas de pesquisa distintas do processamento de linguagem natural, como visão computacional (CAO; FEI-FEI, 2007; WANG; GRIMSON, 2007), bioinformática (LIU et al., 2016; SCHNEIDER et al., 2017), *internet of things* (LIU et al., 2018), entre outras.

A Figura 8 ilustra a esquematização do chamado processo generativo de um documento, que é a base para os modelos de tópicos explorados no presente trabalho. O processo generativo ilustra a maneira como os modelos compreendem a geração de documentos dentro de um *corpus*. No esquema, pode-se notar que os tópicos são representados por distribuições de probabilidade sobre os documentos do *corpus*. Cada tópico é representado por uma mistura de palavras, também representada por uma distribuição sobre o vocabulário. Note que, sucessivamente, o processo generativo determina a escolha de um tópico e em seguida a escolha de um termo do tópico para preenchimento do conteúdo de um dado documento. Todos os modelos de tópicos explorados nas seções a seguir consideram um processo generativo similar ao esquematizado, com variações na forma de determinação das distribuições ilustradas e na forma de escolha de tópicos e palavras para composição dos documentos.

No presente estudo, três modelos de tópicos foram explorados. Apesar de compartilharem semelhanças quanto ao processo generativo citado, tais modelos possuem diferenças arquiteturais significativas. O *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003) é um modelo estatístico que mapeia um *corpus* em um conjunto de tópicos por meio de distribuições probabilísticas, recebendo como entrada um *corpus* codificado em BOW. O *Contextualized Topic Model* (CTM) (BIANCHI; TERRAGNI; HOVY, 2020; BIANCHI et al., 2021) troca a inferência puramente estatística por uma arquitetura de ANNs que aproxima a modelagem probabilística de tópicos, e tem como entrada representações

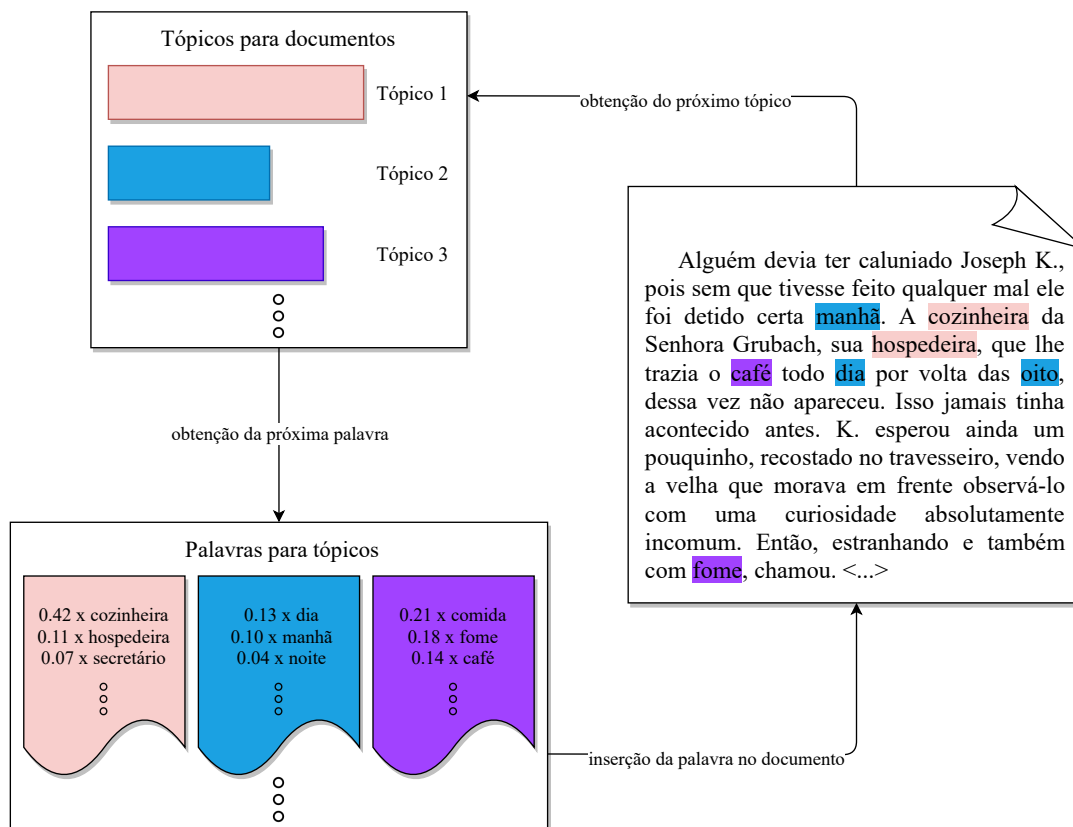


Figura 8 – Processo generativo dos modelos de tópicos estudados. Trecho de *O Processo* de Franz Kafka usado para exemplificar o documento construído (KAFKA, 2019).

de documentos produzidas por meio do BERT em conjunto com a representação BOW. Por sua vez, o *Embedded Topic Model* (ETM) (DIENG; RUIZ; BLEI, 2020) busca unir as características do *word2vec* com a modelagem probabilística do LDA, empregando o primeiro como entrada, utilizando o segundo como base conceitual e usufruindo também de ANNs para realização do processo de inferência.

2.2.1 Latent Dirichlet Allocation

O *Latent Dirichlet Allocation* (LDA) é um modelo probabilístico proposto em (BLEI; NG; JORDAN, 2003). O LDA tornou-se, ao longo dos anos, o modelo de tópicos mais popular, e baseia-se na hipótese de que um processo generativo está por trás da construção de todo e qualquer documento pertencente a um dado *corpus*. O LDA subentende que documentos são compostos de misturas de tópicos, e tópicos são compostos de misturas de palavras. No LDA, cada palavra contribui em maior ou menor grau para a constituição de um tópico, assim como cada tópico contribui em maior ou menor proporção para a constituição de um documento. As proporções de contribuição de cada palavra dentro de um tópico e de cada tópico dentro de um documento são dadas por probabilidades. A partir da hipótese generativa, o LDA trabalha realizando uma espécie de “engenharia

reversa”, inferindo os tópicos que formaram o *corpus* analisado a partir do mesmo.

O processo generativo do LDA pode ser decomposto em três principais etapas segundo (BOYD-GRABER et al., 2017): **geração dos tópicos**, **alocação de tópicos para documentos** e **geração de palavras**. Cada uma dessas etapas é descrita a seguir.

- Na etapa de **geração de tópicos**, o usuário inicialmente deve determinar o número de tópicos distintos no *corpus*. Este valor é dado por K , um hiperparâmetro do modelo. Cada um dos tópicos será composto por meio de uma distribuição Dirichlet $\phi_k \sim Dir(\beta)$, onde o parâmetro de concentração β também é um hiperparâmetro de entrada. A distribuição ϕ_k conterà pesos – ou seja, probabilidades – para cada um dos termos do vocabulário. Contudo, deve-se notar que para fins de análise do modelo, seja ela qualitativa ou quantitativa, normalmente apenas os termos com maior probabilidade dentro de um dado tópico são considerados. Comumente, os dez primeiros termos em valor de probabilidade são utilizados para tal;
- Na **alocação de tópicos** deve-se determinar quais tópicos formam os documentos do *corpus*. Sendo assim, nesta etapa, a ideia é determinar a distribuição de tópicos para documentos. Para cada documento, uma distribuição sobre todos os tópicos determina o conteúdo do mesmo. A distribuição $\theta_d \sim Dir(\alpha)$ definirá as alocações de tópicos entre os diversos documentos, onde α é um hiperparâmetro de entrada que controla a esparsidade da distribuição – isto é, se os documentos serão formados por poucos ou por muitos tópicos. Note que os tópicos com maiores probabilidades na distribuição gerada terão maior presença na formação dos documentos;
- Dado que sabe-se a constituição temática de cada documento, a etapa de **geração de palavras** pode ser iniciada. Nesta etapa, assume-se que cada documento d é composto por N_d palavras. Para cada palavra n do documento d , realiza-se uma atribuição de tópico $z_{d,n} \sim Discrete(\theta_d)$. O tópico irá definir qual palavra será selecionada para compor o documento, pois define a distribuição sobre palavras que será empregada para a escolha. Por fim, escolhe-se a palavra $w_{d,n} \sim \phi_{z_{d,n}}$, utilizando-se a atribuição de tópico. Novamente, deve-se notar que os termos com maior probabilidade na distribuição serão escolhidos com maior frequência para formarem os documentos.

Naturalmente, percebe-se que a hipótese generativa não representa fielmente a forma como um escritor redige um texto, qualquer que seja o assunto do mesmo. O processo aqui descrito tem valia no sentido de que, assumindo tal método generativo, pode-se realizar o procedimento inverso: a partir dos documentos originais, obter os tópicos latentes que constituem os temas ali presentes. O processo de revelar os tópicos a partir do *corpus* é chamado de inferência de tópicos. A inferência consiste em definir os tópicos

latentes que deram origem aos diferentes documentos presentes no *corpus*. Dessa maneira, o problema que deve ser resolvido neste cenário é o de obter a distribuição posterior de variáveis ocultas para cada um dos documentos do *corpus*. Como tal distribuição posterior é intratável, métodos aproximativos são necessários. Na literatura, diversas abordagens para inferência posterior foram utilizadas, como inferência variacional (BLEI; NG; JORDAN, 2003) e amostragem de Gibbs (GRIFFITHS; STEYVERS, 2004).

Em (ASUNCION et al., 2012), é mostrado que a abordagem usando *batches* do *Variational Bayes* (VB) para LDA tem requisitos de memória constantes e converge mais rapidamente que a amostragem de Gibbs colapsada em *batches*. Contudo, a abordagem VB ainda requer uma varredura completa do *corpus* a cada iteração do algoritmo. (HOFFMAN; BACH; BLEI, 2010) propõem uma versão *online* do VB para LDA capaz de receber dados conforme eles chegam, possibilitando a acomodação de conjuntos de dados com enormes proporções. Dessa forma, o *online* VB mostra-se uma alternativa rápida ao VB no contexto de grandes *datasets*, que por vezes não cabem em memória por completo. Os autores mostram que o algoritmo *online* VB é tão simples quanto o VB original, converge mais rapidamente que este em *corpora* amplos e gera tópicos com qualidade equivalente ou superior à sua versão baseada em *batches*.

No aprendizado de máquina, deve-se considerar diversos parâmetros a serem definidos antes da realização do treinamento desejado. Tais parâmetros são chamadas de hiperparâmetros, e são responsáveis por realizar um ajuste fino dos modelos treinados. Desta forma, os hiperparâmetros proporcionam uma melhor adaptatividade dos modelos de aprendizagem, em diferentes contextos de uso. O modelo LDA possui três hiperparâmetros principais: o número de tópicos K e os parâmetros das distribuições Dirichlet, α e β . O número K de tópicos representa o número de temas distintos presentes no *corpus*. Tal valor pode ser oriundo de observação do *corpus* avaliado ou determinado a partir de variação de hiperparâmetros e análise empírica. Note que, em um *corpus* com poucos assuntos, um valor elevado para K produzirá tópicos fragmentados que compartilham semântica parecida. Já em um *corpus* com muitos assuntos, um valor baixo para K levará a tópicos sem significado claro, que agregarão ideias díspares.

Os hiperparâmetros restantes, α e β , modelam as distribuições Dirichlet de tópicos para documentos e de palavras para tópicos empregadas no processo generativo, respectivamente. Estes são chamados de parâmetros de concentração, pois definem a esparsidade das distribuições de tópicos para documentos e de palavras para tópicos. No contexto de α , valores baixos para este parâmetro incutirão em misturas menores de tópicos para documentos. Dessa forma, os documentos do *corpus* serão formados apenas por uma pequena fração dos tópicos existentes. Caso o valor de α seja alto, os documentos serão considerados como misturas de múltiplos tópicos. Sendo assim, um maior número de tópicos irá contribuir para a geração de um documento. No contexto de β , sua variação

irá acarretar na ampliação ou redução da esparsidade na distribuição de palavras para tópicos. Valores baixos farão com que tópicos sejam percebidos como uma mistura de poucas proeminentes palavras. Enquanto isso, valores altos acarretarão na inferência de que múltiplos termos contribuem para a formação de um tópico.

Os hiperparâmetros de concentração do LDA são difíceis de estimar a olho nu durante uma análise subjetiva do *corpus*. Contudo, assim como no caso de K , é possível realizar a variação de hiperparâmetros e procurar o modelo com as configurações que melhor se encaixam na tarefa de modelagem em questão. Idealmente, tal variação de hiperparâmetros é realizada por metodologias como o *grid search*, onde os melhores hiperparâmetros são definidos por meio de uma busca exaustiva no espaço de valores disponíveis e escolhidos mediante validação cruzada (LUTINS, 2017).

2.2.2 Contextualized Topic Model

Contextualized Topic Model (CTM) (BIANCHI; TERRAGNI; HOVY, 2020; BIANCHI et al., 2021) é um modelo de tópicos baseado em ANNs. Sua proposta é a de possibilitar a inferência de tópicos multilíngue a partir de *embeddings* BERT pré-treinados. A ideia por trás do CTM é que mesmo que o modelo seja treinado a partir de um *embedding* representando uma única linguagem, seja possível realizar inferência de tópicos em documentos desconhecidos escritos em outros idiomas. Nesse caso, os tópicos inferidos serão constituídos no idioma de treinamento original. Portanto, o CTM caracteriza-se como um exemplo da abordagem *one-shot learning* dentro de aprendizado de máquina. Nesta abordagem, durante os testes de um modelo, exemplos de uma categoria desconhecida em tempo de treinamento são apresentados e o modelo deve categorizar tais exemplos de forma adequada.

O CTM surgiu da motivação em estender o modelo *Neural-ProdLDA* proposto por (SRIVASTAVA; SUTTON, 2017), modificando a entrada de dados em formato BOW deste por uma entrada que combina *embeddings* BERT e representação BOW ou utiliza apenas BERT, sendo possível optar por ambas as combinações. Dessa forma, informações de contextualização são transferidas ao modelo de tópicos treinado, possibilitando a extração mais efetiva de tópicos do *corpus* estudado. No CTM, a representação SBERT, uma variante do BERT tradicional, é empregada como entrada. Existem duas variações do modelo: uma que recebe apenas os *embeddings* SBERT como entrada – chamada ZeroShotTM – e outra que trabalha com uma entrada composta pelo BOW dos documentos e dos *embeddings* SBERT – chamada CombinedTM. Segundo os autores, a versão ZeroShotTM é indicada para modelagem de tópicos com inferência multilíngue. Já a versão CombinedTM é indicada para modelagem de tópicos comum e foi empregada no presente estudo. Dadas as limitações relativas ao tamanho de sentenças compartilhadas pelos modelos BERT e SBERT, o modelo CTM é indicado para *corpus* que contenham documentos de tamanho

pequeno ou médio. Isto não representa um problema no âmbito do presente estudo, já que postagens em mídias sociais costumam ter um tamanho significativamente inferior à métodos mais tradicionais de comunicação escrita, como cartas ou livros.

Sendo uma extensão ao *Neural-ProdLDA*, o CTM compartilha de arquitetura de funcionamento semelhante. A principal similaridade entre ambos os modelos é no uso do *autoencoder* variacional. Um *autoencoder* variacional (VAE) é uma rede neural utilizada para obter uma representação latente em menor dimensão de um dado conjunto de entradas de forma que um processo generativo seja viabilizado a partir de um conjunto original de dados (ROCCA, 2019). Conseqüentemente, a arquitetura de um VAE encaixa-se dentro do contexto da modelagem de tópicos, já que esta tarefa consiste em encontrar uma representação em menores dimensões – os tópicos – de um *corpus* de origem, de forma que o processo generativo do *corpus* possa ser representado. *Neural-ProdLDA* e CTM são centralizados em um *autoencoder* variacional responsável pelo treinamento de sua rede de inferência. No caso do CTM, a entrada em SBERT é codificada pela rede neural em uma representação latente contínua. Em seguida, o decodificador tem a responsabilidade de mapear a representação num novo conjunto de dados, de forma que as principais *features* da entrada não se percam neste procedimento. Com exceção do formato de entrada, as demais descrições arquiteturais aqui realizadas valem para ambos os modelos.

O CTM assume um processo generativo semelhante ao LDA, considerando que tópicos são compostos por uma combinação de palavras e que documentos são compostos por uma combinação de tópicos. Contudo, as divergências explicitam-se na forma como as distribuições são combinadas para geração de tópicos e documentos e também na maneira como o CTM realiza a inferência das distribuições envolvidas (SRIVASTAVA; SUTTON, 2017). Enquanto no LDA considera-se que as palavras de um documento são formadas pela “mistura” de termos de diversos tópicos, no CTM um *product of experts* com pesos é utilizado para determinar os termos que formam um texto. O *product of experts* (HINTON, 2002) é uma técnica de modelagem de distribuições de probabilidade onde combina-se o resultado de diversas distribuições mais simples. A combinação é realizada por meio do produto entre as funções de densidade das distribuições. Em termos de comparação, pode-se dizer que o *product of experts* se assemelha a uma operação AND entre as distribuições envolvidas, enquanto o modelo de *mixture of experts* – empregado no LDA – assemelha-se a uma operação OR entre as diversas distribuições multinomiais envolvidas, pois realiza uma soma das funções de densidade. Para compreender-se as diferenças entre os modelos de mistura e de produto, deve-se considerar que o modelo de mistura dará probabilidade alta para um evento *e* caso um dos *experts* assumam probabilidade alta para este evento. Por outro lado, no modelo de produto um evento *e* terá probabilidade alta apenas caso nenhum dos *experts* tenha assumido uma probabilidade baixa para o mesmo. Desta forma, percebe-se que enquanto no modelo de mistura um *expert* terá o poder de atribuir probabilidade alta a um dado evento, no modelo de produto será necessário que todos os *experts* contribuam

para tal (WELLING, 2007). Desta forma, utilizando tal modelo pode-se realizar predições mais precisas sobre os documentos em estudo. Isso se contrapõe ao que pode ser percebido no LDA, onde acontece o aparecimento de tópicos de baixa qualidade em razão do modelo não conseguir realizar predições mais nítidas que os tópicos sendo “misturados”. Ademais, as três principais etapas do processo generativo descritas para o LDA continuam valendo para o CTM, excetuando-se as distinções relativas à combinação das distribuições citadas.

Conceitualmente, um dos grandes diferenciais dos modelos CTM em relação aos modelos LDA está na forma de realização da inferência das distribuições que regem a formação dos documentos. Enquanto no LDA métodos de aproximação da distribuição posterior tradicionais são aplicados para inferir as distribuições Dirichlet desejadas, no CTM uma rede neural é utilizada como um aproximador destas distribuições. Mais especificamente, a rede neural em questão é um *autoencoder* variacional (VAE) composto de duas sub-redes: um *encoder* e um *decoder* (SRIVASTAVA; SUTTON, 2017). A rede de *encoding* do VAE é responsável por aprender a representação latente em menores dimensões, e posteriormente, a rede de *decoding* é responsável por gerar a partir da representação latente um conjunto de novos dados, com a mesma dimensão dos dados originais. Os novos dados compartilharão características dos originais, mas serão distintos. No CTM, o VAE é responsável por inferir aproximações das distribuições de documentos para tópicos e tópicos para palavras. As aproximações das distribuições Dirichlet são realizadas por meio da aproximação de Laplace. Ainda, para otimização da função objetivo variacional, determinada a partir das aproximações, é empregado o uso de gradiente descendente estocástico.

Em relação aos hiperparâmetros, no CTM, ao contrário do LDA, os *priors* das distribuições de tópicos e documentos são parâmetros aprendidos pelo próprio modelo durante o treinamento (BIANCHI; TERRAGNI; HOVY, 2020). Desta forma, o principal hiperparâmetro configurável no CTM é o número de tópicos K . Existem outros hiperparâmetros relacionados à arquitetura da rede neural a ser utilizada no modelo e que controlam a taxa de aprendizado do mesmo. No presente trabalho, tais hiperparâmetros foram utilizados em suas configurações padrão de acordo com o disponibilizado originalmente pelos autores do modelo em sua implementação. Ainda, pode-se citar que o CTM pode utilizar o LDA como método de modelagem, ao contrário do *ProdLDA* usado por padrão e também empregado no atual trabalho.

2.2.3 *Embedded Topic Model*

Embedded Topic Model (ETM) foi proposto em (DIENG; RUIZ; BLEI, 2020) como a união entre a modelagem probabilística do LDA e a riqueza semântica dos *word embeddings* – mais especificamente, do *word2vec*. A ideia no ETM é que tópicos sejam representados por pontos no espaço vetorial de *word embeddings*. Dessa forma, tópicos serão dispostos

no espaço vetorial de acordo com a similaridade compartilhada com os demais *word embeddings* em sua vizinhança. Além de considerar que tópicos pertencem ao espaço vetorial de *embeddings*, o ETM mantém as características de modelagem probabilística do LDA. No ETM, documentos e tópicos também são definidos por distribuições de probabilidade – no caso dos tópicos, uma distribuição que leva em consideração a semelhança entre os *embeddings* de termos e do tópico. O *corpus* de entrada no ETM é codificado utilizando *word2vec*.

Além disso, o ETM foi criado para ser uma alternativa melhor ao LDA no contexto de extração de tópicos usando *corpora* e vocabulários muito extensos. O LDA depende de um pré-processamento significativo quanto à remoção de termos mais e menos frequentes do *corpus* para gerar bons tópicos. Tal fator pode ser problemático no caso de treinamentos envolvendo vocabulários extensos, já que uma enorme quantidade de termos deverá ser eliminada antes do treinamento e termos com grande importância para o *corpus* analisado podem ser perdidos em razão disso (DIENG; RUIZ; BLEI, 2020). Segundo seus autores, enquanto o LDA decai em qualidade à medida que o vocabulário do *corpus* de treinamento aumenta, o ETM mantém bom desempenho. Além dos pontos citados, as principais diferenças entre o ETM e o LDA concentram-se no contexto da distribuição de tópicos sobre termos e também no método de inferência empregado.

Assim como no caso do modelo CTM, o modelo ETM carrega muitas semelhanças com o LDA em seu processo generativo. Aqui, documentos também são formados por misturas de tópicos. Contudo, tópicos no ETM não são simplesmente distribuições sobre termos, mas sim vetores no espaço vetorial de *word embeddings*. O ETM emprega um modelo que obtém o produto interno da matriz de *word embeddings* pelo *embedding* do tópico. De acordo com a semelhança entre os vetores de palavras e tópicos, as probabilidades são dadas. Caso os vetores de um conjunto de palavras carreguem muitas semelhanças com o vetor de um tópico, então tais termos terão probabilidade alta de pertencerem ao tópico em questão (DIENG; RUIZ; BLEI, 2020).

A aplicação de uma função *softmax* garante que a obtenção de uma palavra na formação de um documento será realizada via amostragem a partir de uma distribuição de probabilidade. Enquanto no modelo CBOW um conjunto de palavras era utilizado como um contexto para determinar a palavra-alvo a ser aprendida, no ETM o *embedding* do tópico é usado como o vetor de contexto. Contudo, no ETM as palavras são amostradas em um contexto de documentos ao invés do contexto de janela de palavras, como o empregado no *word2vec*.

Uma diferença substancial no processo generativo do ETM está nas distribuições empregadas. Além do emprego de *softmax* para determinar a distribuição de tópicos sobre palavras, o ETM emprega a distribuição logístico-normal para amostragem da distribuição de documentos para tópicos. Tal função foi utilizada para facilitar a reparametrização a

ser realizada pelo algoritmo de inferência.

O ETM emprega inferência variacional e inferência amortizada para aproximação da posterior intratável do problema. Portanto, o problema da inferência é reimaginado como um problema de otimização. O objetivo aqui, assim como nos demais modelos de tópicos apresentados que seguem o método VB, é maximizar o *Evidence Lower Bound* (ELBO) do modelo. Uma rede neural é empregada como rede de inferência para determinação de alguns dos parâmetros necessários à otimização, mas diferentemente do modelo CTM, a rede não é responsável pela realização de todo o procedimento de aproximação. As aproximações para a inferência são realizadas de forma iterada, e subamostragem de dados é aplicada para otimizar o processamento de um *corpus* extenso de documentos. Esta técnica consiste em realizar as aproximações apenas para cada subconjunto dos dados, evitando o processamento de todo o conjunto de uma única vez.

Como nos demais modelos apresentados, o número K de tópicos a serem identificados no *corpus* é o principal hiperparâmetro do modelo ETM. Demais hiperparâmetros relativos ao treinamento e arquitetura do modelo são definidos no pacote do modelo. Para o presente trabalho, as configurações padrões para os hiperparâmetros existentes no modelo disponibilizado pelos autores originais foram empregadas.

2.3 Análise textual baseada em categorias lexicais

A linguagem humana carrega muitas informações sobre os indivíduos que a utilizam, já que pessoas distintas comunicam-se de forma diferente nos mais diversos âmbitos sociais (TAUSCZIK; PENNEBAKER, 2010). A categorização e análise sistemática das palavras usadas pelos seres humanos em seu dia-a-dia tornou-se mais palatável com o advento dos computadores modernos. Além disso, o surgimento da Internet e de novos modelos estatísticos para representar a linguagem também ampliaram esses horizontes. No presente estudo, ferramentas para análise textual automatizada foram exploradas em adição ao uso da modelagem de tópicos, com o objetivo de extrair a semântica latente das postagens oriundas do *Reddit*. Mais especificamente, as ferramentas foram utilizadas para extração das categorias lexicais associadas aos tópicos latentes obtidos por meio da modelagem de tópicos. Para tanto, duas ferramentas apropriadas para a tarefa foram empregadas: *Linguistic Inquiry and Word Count* e *Empath*.

O *Linguistic Inquiry and Word Count* (LIWC), pronunciado “*Luke*”, é uma ferramenta de análise textual que contabiliza palavras em categorias psicologicamente significantes. O LIWC surgiu das dificuldades impostas pela tarefa de análise textual realizada de forma manual, utilizando julgadores humanos para determinar a semântica associada a conjuntos textuais diversos (TAUSCZIK; PENNEBAKER, 2010). Comumente, juízes humanos discordam de forma significativa sobre a semântica associada a um determinado

texto, o que dificulta a concordância sobre a análise. Além disso, a avaliação humana de conjuntos textuais amplos torna-se mais custosa e complexa à medida que o *corpus* é ampliado. Finalmente, os autores perceberam que juízes humanos também são emocionalmente impactados pelo conteúdo lido. Nesse contexto, surgiu a ideia para um programa que automatize a contagem de palavras em categorias psicológicas, que veio a ser o LIWC. A validação das categorizações oriundas do LIWC também passou pelo crivo do julgamento humano de forma a determinar a qualidade das associações.

O LIWC é uma ferramenta que tem o poder de processar textos dados como entrada e categorizar as palavras pertencentes ao mesmo em categorias pré-definidas. A categorização é realizada por meio de um dicionário interno à ferramenta, responsável por mapear palavras em categorias. Em sua versão de 2007, existem mais de sessenta categorias de palavras LIWC e o dicionário da ferramenta foi traduzido para diversas linguagens em suas diferentes versões, incluindo o português (FILHO; PARDO; ALUÍSIO, 2013). O LIWC possui categorias para processos linguísticos, processos psicológicos, preocupações pessoais, entre outras. Por exemplo, dentro dos processos psicológicos, existem as categorias associadas a processos afetivos: emoção positiva – abreviada de *posemo* e emoção negativa – abreviada de *negemo*. Dentro de *posemo*, pode-se citar termos como “love” ou “nice”, enquanto dentro de *negemo* pode-se mencionar os termos “hurt” e “ugly”. Resultados mostram que o LIWC apresenta alta correlação com o julgamento humano (PENNEBAKER et al., 2007).

A ferramenta LIWC foi empregada em diversos estudos como forma de identificação de categorias lexicais presentes em textos das mais variadas fontes (RAMIREZ-ESPARZA et al., 2008; EICHSTAEDT et al., 2018; TADESSE et al., 2019).

Por sua vez, o *Empath* é uma ferramenta de análise textual baseada em aprendizado profundo. Essa ferramenta tem como foco representar um conjunto léxico vivo, em frequente aprimoramento e evolução, se comparado às abordagens estáticas focadas em listas de palavras, como é o caso do LIWC (FAST; CHEN; BERNSTEIN, 2016). O *Empath* permite que, além das categorias pré-definidas inclusas no mesmo, novas categorias sejam definidas por meio de termos chamados de “semente”: palavras com semântica associada que em conjunto definem uma categoria. *Empath* é treinado por meio de aprendizado profundo em mais de 1 bilhão de palavras de ficção moderna, usando uma arquitetura baseada no modelo *skipgram*. O objetivo da arquitetura é proporcionar o aprendizado de associações entre palavras e seu contexto de uso. Em seguida, um vocabulário de mais de cinquenta mil palavras é mapeado para as duzentas categorias-padrão da ferramenta, por meio do emprego de métricas de similaridade no espaço vetorial produzido. O emprego de tal arquitetura é a razão da viabilidade da geração de novas categorias lexicais proporcionada pelo *Empath*. Além disso, as categorias do *Empath* apresentam alta correlação com categorias semelhantes no LIWC. Entretanto, o *Empath* é treinado para categorizar palavras escritas apenas em

inglês.

2.4 Trabalhos relacionados

A exploração dos ambientes das redes sociais com modelagem de tópicos já foi realizada múltiplas vezes na literatura, de diferentes formas. Dentro do âmbito de saúde mental, problemas como depressão, ansiedade e tendências suicidas apresentados por meio de linguagem nas redes são alguns dos pontos de interesse de pesquisadores da área.

Por exemplo, o estudo de (EICHSTAEDT et al., 2018) procurou realizar a identificação da depressão em usuários do *Facebook* que visitaram uma unidade de emergência médica específica a partir de suas postagens na rede. O grupo de pacientes do hospital foi separado em dois: um grupo de controle, sem diagnóstico prévio de depressão, e um grupo de pacientes com diagnóstico de depressão em seus registros médicos. Foram utilizados como *features* de entrada para apoiar a decisão de um classificador neste estudo o conteúdo, tamanho, frequência e padrões temporais de postagens. Um dos achados do estudo é que *features* de linguagem, como tópicos, contribuem melhor para a predição de depressão do que *features* como tamanho, frequência ou período temporal de postagens. Os tópicos obtidos via LDA foram analisados também com os dicionários de palavras oriundos do LIWC. A partir da análise, os autores notaram tópicos sensíveis ao diagnóstico da depressão, como solidão, hostilidade, ruminação, ansiedade e queixas somáticas. Por fim, notou-se que predições realizadas pelo classificador treinado com *features* de linguagem tiveram taxa de sucesso semelhante à de pesquisas de triagem de possíveis enfermos.

Ainda no âmbito de predição da depressão, (TADESSE et al., 2019) examinou postagens de usuários no *Reddit* em busca de fatores que revelassem traços de depressão nos membros da rede. No estudo, um conjunto de termos comumente associado a usuários depressivos foi descoberto por meio de um conjunto de técnicas que incluem tópicos obtidos via LDA, dicionário LIWC e *features n-gram*. Os três métodos foram empregados para extração de *features* a partir do *corpus* pré-processado do *Reddit*, com o intuito de alimentar classificadores. Por sua vez, os classificadores ficam responsáveis por determinar a tendência à depressão entre os usuários. No contexto da extração de tópicos usando LDA, pode-se citar que os autores notaram a qualidade dos tópicos gerados, englobando temas como depressão, problemas emocionais, agressividade e solidão, entre outros.

Em (RAMIREZ-ESPARZA et al., 2008), um estudo das postagens em fóruns de depressão nas línguas inglesa e espanhola foi realizado. O estudo dividiu-se em duas etapas: na primeira, postagens em fóruns de depressão e de câncer de mama foram coletadas em inglês e espanhol para que seu conteúdo fosse avaliado utilizando LIWC. As postagens de usuários do fórum sobre câncer formaram o grupo de controle do estudo. Percebeu-se que os usuários depressivos utilizavam a linguagem de forma similar, mesmo

com as devidas distinções entre ambas as linguagens, empregando com maior frequência termos relacionados a negatividade e pronomes em primeira pessoa do singular – fatores constatados como traços depressivos na linguagem (AL-MOSAIWI; JOHNSTONE, 2018). A segunda etapa consistiu da coleta de um outro conjunto de dados, formado também por postagens em inglês e espanhol em fóruns de depressão, desta vez para extração dos tópicos falados durante as discussões. Um método baseado em análise de fatores foi empregado, produzindo tópicos a partir do *corpus* de forma similar ao LDA. Temas semelhantes surgiram entre os falantes de ambos os idiomas, mas percebeu-se que alguns tópicos no inglês tendiam a preocupações medicinais, enquanto alguns tópicos em espanhol tendiam a preocupações sociais.

Indivíduos procuram redes sociais para buscar conselhos e orientações sobre outros distúrbios de saúde mental, e não apenas a depressão. Muitos fóruns e redes sociais contêm grupos voltados para a discussão de transtornos específicos. O *Reddit* possui diversas comunidades temáticas sobre transtornos mentais como o *r/depression*, *r/bipolar*, entre outros. Desta forma, a identificação de múltiplos distúrbios utilizando postagens em redes sociais pode ser viabilizada. Contudo, o conteúdo existente nestas comunidades não representa necessariamente distúrbios descritos no manual diagnóstico e estatístico de desordens mentais (em inglês, DSM) (ASSOCIATION et al., 2013) – o padrão *de facto* para categorização de desordens mentais por profissionais médicos. Em (GAUR; KURSUNCU; ALAMBO, 2018), os autores empregam técnicas de classificação em múltiplas categorias para mapear *subreddits* em categorias do DSM-V. Desta forma, profissionais de saúde mental podem ter a seu dispor uma ferramenta mais precisa para análise do histórico de saúde mental de seus pacientes, englobando dados oriundos de mídias sociais devidamente categorizados. LDA, *word embeddings* e bases de conhecimento médico curado foram utilizados para realização da categorização no estudo.

Ainda, a análise de evolução do tom emocional das postagens de usuários com transtornos mentais em redes sociais também já foi explorada na literatura. Em (SILVEIRA; SILVA; MURAI, 2020) os autores determinam através de análise de sentimentos o tom emocional de postagens no *Reddit* em sub-fóruns sobre suicídio, ansiedade, depressão e bipolaridade. Utilizando as árvores de discussão na rede, determinou-se que o tom emocional de postagens do autor de uma submissão varia positivamente à medida que o mesmo interage com comentários de tom emocional positivo na mesma *thread* – como é chamada uma discussão na rede. A evolução positiva do tom emocional também foi notada ao comparar a postagem inicial – a submissão – do usuário com sua última postagem na *thread* relativa. Ainda, o trabalho buscou realizar a predição do tom emocional do último comentário de um usuário dada sua postagem inicial e demais comentários da respectiva *thread*. Para tal, as submissões e comentários foram codificados por meio de *word embeddings* e alimentaram um classificador *multilayer perceptron* responsável por prever o tom emocional do último comentário do autor da submissão. O modelo proposto previu

com boa acurácia a evolução do tom emocional, reforçando a ideia de que o contexto de discussão da *thread* e sua evolução podem contribuir para uma evolução positiva do quadro emocional do usuário acometido por transtornos mentais. Os autores prosseguem propondo que o modelo projetado pode auxiliar intervenções de profissionais de saúde em discussões na rede. Isto poderia ser realizado em casos onde o tom emocional do indivíduo não evolui positivamente em cenários extremos, como discussões sobre suicídio, por exemplo.

Os trabalhos citados acima buscam identificar traços de depressão e de demais distúrbios mentais em postagens de redes sociais. Contudo, todos os trabalhos citados analisam *corpora* em inglês ou espanhol para extração de tópicos ou análise de sentimentos. Pode-se destacar como contribuição do presente estudo a busca pela identificação de temáticas relacionadas à depressão no *Reddit* em língua portuguesa. Além disso, propõe-se uma análise de tópicos associados à depressão no contexto de postagens em português e em inglês no *Reddit*, com o intuito de identificar semelhanças e distinções entre os temas latentes em ambas as línguas. Uma contribuição adicional do presente trabalho é a comparação de resultados obtidos com o uso de diferentes modelos para extração de tópicos empregando um mesmo *dataset* de origem, utilizando desta forma diferentes arquiteturas para modelagem de tópicos. Pode-se destacar tal fator como uma diferença positiva em relação aos trabalhos citados anteriormente, já que os mesmos focavam seu estudo em apenas um tipo de arquitetura ou modelo para a realização de tarefas similares.

3 Uma Metodologia para Análise de Tópicos de Depressão no Reddit

Neste capítulo, a solução do presente trabalho é apresentada e discutida. As diversas etapas do estudo são detalhadas, abordando desde a coleta e construção da base de dados usada, passando pela preparação dos dados e treinamento dos modelos de tópicos, chegando até a formulação da estratégia de análise a ser seguida para avaliar a proposta. A Figura 9 resume em alto nível as fases do estudo discutidas neste capítulo. Os códigos implementados para a realização desta monografia foram escritos em linguagem Python (ROSSUM; JR, 1995), e fazem uso de diversas bibliotecas de NLP e aprendizado de máquina nomeadas quando relevante. Além disso, os repositórios de código associados às fases do estudo descritas abaixo também são indicados.

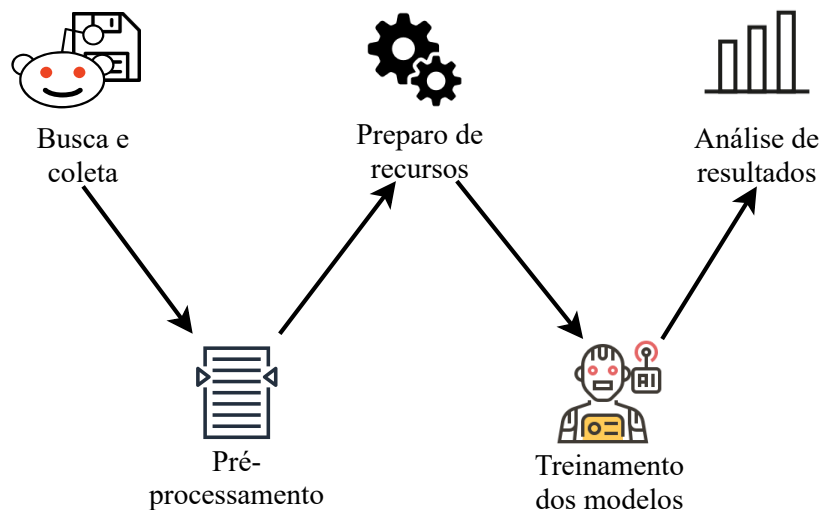


Figura 9 – Fases de construção do estudo realizado.

3.1 Construção da base de dados

Reddit

A presente monografia tem como foco os relatos de usuários na rede social *Reddit*. O *Reddit* é uma rede social criada em 2005 (CRUNCHBASE, 2021) e baseada em um sistema de votação que permite aos seus usuários fornecer e selecionar o que acham relevante ou não a ser exibido na rede (GILBERT, 2013). Usuários podem realizar postagens chamadas de “submissões”, que iniciam uma discussão sobre algum tema em específico definido no *post* ou realizar comentários em discussões pré-existentes – também chamadas

de *threads*. Os comentários também podem ser feitos em resposta a outros comentários numa mesma discussão. O criador de uma submissão pode interagir com os comentários na discussão, impulsionando o diálogo entre os membros. Além disso, a rede tem um sistema de comunidades – chamadas de *subreddits* – onde usuários podem participar de discussões temáticas relativas a seus interesses pessoais. A comunicação no *Reddit* é realizada primordialmente por meio de texto, mas imagens e vídeos também são permitidos no caso de submissões. Comumente, são utilizados *links* para páginas externas com artigos, imagens ou vídeos. Contudo, o discurso textual é usado de forma exclusiva na grande maioria das discussões temáticas dentro da rede, e durante a avaliação manual de parte das postagens relativas ao tema estudado coletadas na rede percebeu-se que a presença de símbolos como *emojis* ou *emoticons* é pequena. Ainda, a plataforma também permite transmissões de vídeo ao vivo. Na data de escrita da presente monografia, o *Reddit* é o décimo-nono sítio *web* mais visitado no mundo¹.

Como dito acima, o *Reddit* é estruturado por meio de um sistema de votação. Cada submissão ou comentário realizado na rede pode receber votos dos demais usuários. Um *upvote*, que seria um voto positivo, torna aquela submissão ou comentário mais relevante do ponto de vista da comunidade ou discussão, respectivamente. Desta forma, submissões com mais *upvotes* aparecem com maior prioridade na listagem de postagens de uma determinada comunidade. Enquanto isso, comentários com mais *upvotes* aparecem em primeiro lugar dentro das discussões a que se referem. No caso de submissões e comentários com muitos *downvotes*, o efeito é exatamente o contrário. Consequentemente, o *Reddit* prioriza a exibição de conteúdos que foram definidos como positivamente relevantes pelos seus usuários.

O *Reddit* é caracterizado principalmente pelo fato de possuir discussões hierarquizadas, onde uma submissão de um usuário pode receber vários comentários em resposta. Os comentários, por si, também podem ser respondidos iterativamente, aprofundando a hierarquia de comentários da discussão. Inicialmente, o objetivo da coleta era obter o conteúdo tanto das submissões que abrem as discussões quanto dos comentários referentes a elas. Contudo, por meio de observação empírica de postagens em português, foi observado que a maioria dos comentários nas submissões analisadas consistiam de respostas curtas, sem menções a situações relativas à depressão, e que normalmente expressavam apoio e suporte ao autor da submissão. Além disso, também foi avaliado empiricamente que o conjunto de dados composto também por comentários de submissões não agregava maior semântica aos tópicos extraídos pelos modelos treinados com tal *corpus* em português. Desta forma, optou-se por utilizar apenas o conteúdo das submissões principais, sem os comentários, para construção de ambas as bases de dados – em português e em inglês – exploradas neste estudo.

¹ <<https://www.alexacom/siteinfo/reddit.com>>

A construção da base de dados a partir do *Reddit* necessitou do uso de *scripts* para coleta e armazenamento das postagens na rede social. A busca por postagens de interesse foi realizada por meio da *API* disponibilizada pela plataforma *Pushshift*². O *Pushshift* é uma plataforma que coleta, armazena e disponibiliza dados provenientes de postagens no *Reddit* (BAUMGARTNER et al., 2020). A *API* possibilita a realização de buscas dentro de *subreddits* específicos, dentro de um período de tempo delimitado e utilizando palavras-chave de busca, funcionalidades que foram extensamente utilizadas para realização das buscas aqui descritas. Realizadas as buscas desejadas, o conteúdo obtido foi salvo em um banco de dados MongoDB no formato JSON³. A etapa de coleta de dados foi realizada parcialmente de forma automatizada por meio de uma aplicação criada na nuvem da Amazon Web Services e também por meio de execução na própria máquina usada para estudo. O código desta etapa encontra-se disponível para acesso na plataforma Github⁴.

Definidas a metodologia e as ferramentas a serem empregadas na coleta de dados, foi necessário definir quais *subreddits* e que conjunto de postagens seriam coletados para composição do conjunto de dados. Para a base de dados em inglês, foi possível escolher o *subreddit depression*, focado no tema de interesse, para a realização da coleta. Neste caso, apenas uma busca por período de tempo foi realizada, já que não foi vista a necessidade de delimitar a busca feita nesta comunidade por meio de palavras-chave, pois, em princípio, o *subreddit* referido agrega apenas postagens que discutem o tema depressão. Um conjunto de mais de 30000 submissões em inglês foi coletado a partir dos critérios de busca. A Tabela 1 detalha as informações relativas à coleta de dados referentes ao *corpus* em inglês.

Por outro lado, as escolhas para a construção da base de dados em português se deram de forma distinta. Em primeiro lugar, não existe uma comunidade em português no *Reddit* que seja equivalente ao *depression* em inglês, ou seja, que aborde exclusivamente o tema. Dessa forma, foi necessário recorrer a outras comunidades para obtenção dos dados. Em língua portuguesa, o *subreddit* intitulado *desabafos* é composto por relatos dos mais variados tipos, o que inclui *desabafos* relativos à depressão. Dessa forma, inicialmente, o *subreddit* foi escolhido para compor a base de dados em português. Contudo, em razão de sua constituição heterogênea, foi preciso definir termos de busca para delimitar o conjunto de postagens da comunidade a ser coletado. Um conjunto de *strings* de busca foi definido para auxiliar nesta busca. As palavras-chave utilizadas foram definidas de forma empírica, já que na prática trouxeram resultados que abarcam discussões sobre depressão. Entretanto, percebeu-se que mesmo com a popularidade da comunidade *desabafos*, poucos relatos associados à depressão estavam sendo obtidos por meio da *API* do *Pushshift*. Sendo assim,

² <<https://pushshift.io>>

³ *JavaScript Object Notation*: formato de dados destinado a representar mapeamentos do tipo chave-valor e tipicamente usado em aplicações *web*.

⁴ <<https://github.com/lffloyd/reddit-posts-gatherer>>

decidiu-se realizar as buscas com o mesmo conjunto de palavras-chave em outro *subreddit* em português: *brasil*. Esta comunidade é generalista e busca agregar os mais diversos assuntos relativos ao Brasil dentro da rede, incluindo discussões sobre política, postagens de humor, entre outros. A escolha da comunidade não se deu apenas pelo grande número de usuários e postagens: notou-se nela a existência de postagens relatando problemas emocionais diversos, algumas das quais diretamente associadas à depressão segundo, seus autores. A Tabela 1 possui informações sobre a coleta de dados realizada em submissões de língua portuguesa no *Reddit*, incluindo *subreddits*, termos de busca e período de postagens. Deve-se notar que as postagens coletadas no presente estudo não representam todas as postagens existentes no *Reddit* nos períodos delimitados, e sim uma amostra das mesmas. Ainda, a Tabela descreve características adicionais de ambos *corpora* explorados nesse estudo, como o número de submissões coletadas, o número de *tokens* ou palavras únicos nos conjuntos de documentos e o número médio de *tokens* por documento.

Tabela 1 – Informações sobre a coleta de submissões em ambos os idiomas.

idioma	<i>subreddits</i>	palavras-chave	período	submissões coletadas	<i>tokens</i> únicos	<i>tokens</i> por documento
inglês	<i>depression</i>	-	2009–2021	32165	151759	236
português	<i>brasil</i> <i>desabafos</i>	“depressão” “suicídio” “diagnóstico depressão” “tratamento depressão”	2008–2021	3404	81451	370

3.2 Limpeza e Pré-processamento de dados

No contexto de modelagem de tópicos utilizando LDA, a limpeza dos dados a serem trabalhados é parte essencial para a extração de tópicos com maior qualidade. Dentro de um *corpus*, algumas informações não são relevantes para a modelagem de tópicos e são comumente removidas antes do treinamento de modelos. Uma etapa de pré-processamento do *corpus* foi realizada, sendo ela a responsável por produzir o vocabulário e realizar a divisão de documentos necessária para o treinamento dos três tipos de modelos estudados. As devidas adaptações foram realizadas para acomodar as diferenças em interfaces de entrada utilizadas pelos modelos. O pré-processamento aqui descrito foi realizado de forma semelhante em ambos os conjuntos de dados construídos na etapa anterior, salvo onde descrito o contrário, e o código referente encontra-se no Github⁵. No repositório, o código referente às demais etapas do estudo – com exceção da coleta de dados – também pode ser visualizado. As etapas de pré-processamento realizadas foram as seguintes:

1. Caracteres *newline* e aspas simples foram removidos dos documentos.

⁵ <<https://github.com/lffloyd/reddit-topic-modelling>>

2. Os documentos foram *tokenizados*, ou seja, quebrados em listas de termos, também chamados *tokens*, a partir das *strings* que representavam os textos por completo. Nesta etapa, acentuações também foram removidas. Para a *tokenização*, a biblioteca para modelagem de tópicos *gensim* foi usada, mais especificamente sua função `simple_preprocess` (ŘEHŮŘEK; SOJKA, 2010).
3. Nesta etapa, é realizada a rotulação de *part-of-speech* (POS) dos termos. Para cada um dos *tokens* dos documentos, foi atribuída uma categoria gramatical como verbo, substantivo, advérbio, entre outras. Além disso, para cada um dos *tokens* foi obtido seu lema, etapa chamada de lematização. A lematização consiste em representar os termos, ou lexemas, dos textos a partir de seus lemas. Por exemplo, dado o termo “fazer”, seu lema será “faz”. Isto vale também para demais categorias de palavras, e não apenas para verbos. A lematização tem como objetivo associar termos com lemas em comum, reduzindo a dimensionalidade do vocabulário trabalhado, mas mantendo sua semântica. Por fim, a rotulação gramatical normalmente funciona melhor se a estrutura original dos textos for preservada, já que determinados *tokens* podem perder a semântica de uso fora de seu contexto, tornando difícil sua desambiguação. Em razão disso, a estrutura original das frases de cada documento foi preservada até este momento, com a exceção de acentos e alguns caracteres removidos. Para realização da rotulação de POS e lematização no presente trabalho foi utilizada a biblioteca *spaCy*, e essas operações são realizadas simultaneamente pela biblioteca (HONNIBAL; MONTANI, 2017). Ao fim desta etapa, cada *token* foi transformado em um objeto contendo três campos: *token*, que indica o *token* original presente no documento; *lema*, que indica o lema associado ao *token*; e POS, que indica a categoria de *part-of-speech* do *token*.
4. Em seguida, a remoção de objetos representando *tokens* de categorias de POS indesejadas foi realizada. A determinação das categorias de POS relevantes para ambos os *corpus* estudados foi feita de forma empírica. A princípio, todas as categorias de termos foram mantidas nesta etapa, mas isto produzia tópicos de baixa qualidade e difícil distinção semântica. Um experimento seguinte consistiu em manter apenas termos categorizados como verbos, adjetivos ou substantivos no *corpus*, tanto em inglês quanto em português. Para o *corpus* em inglês a decisão surtiu efeitos positivos conforme avaliado, já que tópicos de melhor qualidade foram obtidos, definindo a abordagem a ser usada para este idioma. Contudo, o resultado não se repetiu no caso do *corpus* em português, onde os tópicos mantiveram-se de baixa qualidade. Os melhores resultados para o *corpus* em português foram obtidos mediante a manutenção de apenas substantivos como categorias de POS, e esta abordagem foi a escolhida para o idioma no prosseguimento do estudo. Deve-se notar que, ao fim da etapa, os documentos ainda são formados por listas de objetos que contém *token*,

lema e categoria POS.

5. Nesse momento, não existe mais necessidade de armazenamento das categorias POS nos objetos de *tokens*. Os objetos são reduzidos a seus lemas, e desta forma, cada documento passa a ser formado por uma lista de lemas.
6. Nesta etapa, foi realizada a remoção de *stopwords* do *corpus* resultante. *Stopwords* são termos que repetem-se em demasia num *corpus* e possuem pouco significado semântico associado, podendo prejudicar a qualidade dos tópicos extraídos. Exemplos de *stopwords* são artigos e preposições. Aqui, foram utilizadas listas de *stopwords* típicas do português e do inglês pré-definidas pela biblioteca *nltk* (BIRD; KLEIN; LOPER, 2009).
7. Finalmente, documentos que ficaram vazios em razão das etapas anteriores de pré-processamento foram removidos. Ao fim desta etapa, dos documentos originais restaram 3394 documentos no *corpus* em português e 32092 documentos no *corpus* em inglês.

Produção do vocabulário

Ao produzir um vocabulário para modelagem de tópicos, deve-se notar que nem todos os *tokens* presentes no *corpus* serão de especial importância. Dessa forma, deve-se analisar com atenção tanto os *tokens* que ocorrem em demasia – as *stopwords* – quanto os que ocorrem muito pouco. Ambos os tipos de *tokens* são considerados irrelevantes ao realizar modelagem de tópicos. No caso das *stopwords*, sua alta frequência num *corpus* produz ruídos nos tópicos extraídos do mesmo, diminuindo sua riqueza semântica. Listas de *stopwords* como as utilizadas na etapa anterior possuem palavras frequentes típicas, mas naturalmente não consideram as *stopwords* específicas do *corpus* trabalhado. Por outro lado, os *tokens* que aparecem muito pouco também são indesejáveis, pois ocorrem tão raramente que acabam por contribuir muito pouco no âmbito da formação de um tópico. Sendo assim, agregam pouca informação aos tópicos. Portanto, a determinação das faixas de frequência em documentos para remoção desses tipos de *tokens* é um passo fundamental na construção de um vocabulário adequado à tarefa. Existem diversas técnicas para contabilizar a frequência de termos em um *corpus*, como frequência em documentos (FD), informação mútua, frequência do termo–inverso da frequência nos documentos, entre outras (YANG; PEDERSEN, 1997; RAMOS et al., 2003).

O vocabulário do *corpus* estudado passou por uma etapa de remoção de *stopwords* e termos pouco frequentes. Em razão das especificidades de cada *corpus*, essa etapa foi realizada por meio de avaliação empírica das estatísticas de frequência do vocabulário formado em português e em inglês. A técnica de FD foi empregada para determinação da frequência de termos no *corpus*. Essa técnica consiste em contabilizar em quantos

documentos cada *token* ocorreu. Para realização da análise, histogramas detalhando as faixas de frequência em documentos foram gerados e as faixas a serem removidas foram definidas a partir de sua observação. Os parágrafos a seguir detalham o processo para o *corpus* de cada uma das linguagens. Em seguida, a classe **Dictionary** da biblioteca *gensim* foi utilizada, realizando a filtragem dos termos indesejados do *corpus*. Além de armazenar o vocabulário, o dicionário tem a função de mapear *tokens* para identificadores numéricos.

A Tabela 2 mostra as faixas de frequência em documentos e a quantidade de *tokens* únicos existente em cada uma delas para o *corpus* em português. Nesse momento, o vocabulário possuía 11499 *tokens* distintos. Na tabela, pode-se notar que a faixa entre 0% e 10% concentra quase a totalidade dos *tokens*. Isto indica que a extrema maioria das palavras ocorre em um conjunto muito pequeno de documentos. Percebe-se também que não existem *tokens* com alta presença no *corpus*, já que as faixas superiores de frequência apresentam contagens baixas.

Tabela 2 – Faixas de frequência em documentos e número de *tokens* únicos existentes para o *corpus* em português.

Porcentagem dos documentos	Número de <i>tokens</i>
De 0% até 10%	11429
Acima de 10% até 20%	44
Acima de 20% até 30%	15
Acima de 30% até 40%	5
Acima de 40% até 50%	1
Acima de 50% até 60%	4
Acima de 60% até 70%	1
Acima de 70% até 80%	0
Acima de 80% até 90%	0
Acima de 90% até 100%	0

A partir dessas observações, foi criado um histograma concentrado no intervalo de interesse, referente à frequência entre 0% e 10% dos documentos, de forma a determinar o valor mínimo de frequência em documentos a ser utilizado para o *corpus*. A Figura 10 ilustra o histograma gerado. Observando o histograma, percebe-se que a maioria dos *tokens* está concentrada na faixa entre 0% e 1% de frequência. Diante disso, optou-se pela remoção de *tokens* do vocabulário cuja frequência fosse inferior a 1%. Nenhuma limitação superior de frequência foi empregada, já que não existem palavras predominantes no *corpus*. Dessa forma, apenas *tokens* que apareceram em menos de 33 documentos foram removidos. Ao fim desta etapa, o vocabulário foi salvo, passando a ter apenas $|V| = 872$ *tokens*.

Para o conjunto de dados em inglês, a Tabela 3 mostra as faixas de frequência em documentos e a quantidade de *tokens* únicos existente. Originalmente, existiam 29117 *tokens* únicos no vocabulário. Similarmente ao observado no caso do *corpus* em português, a faixa entre 0% e 10% dos documentos foi a que mais concentrou *tokens*. Também pode-se

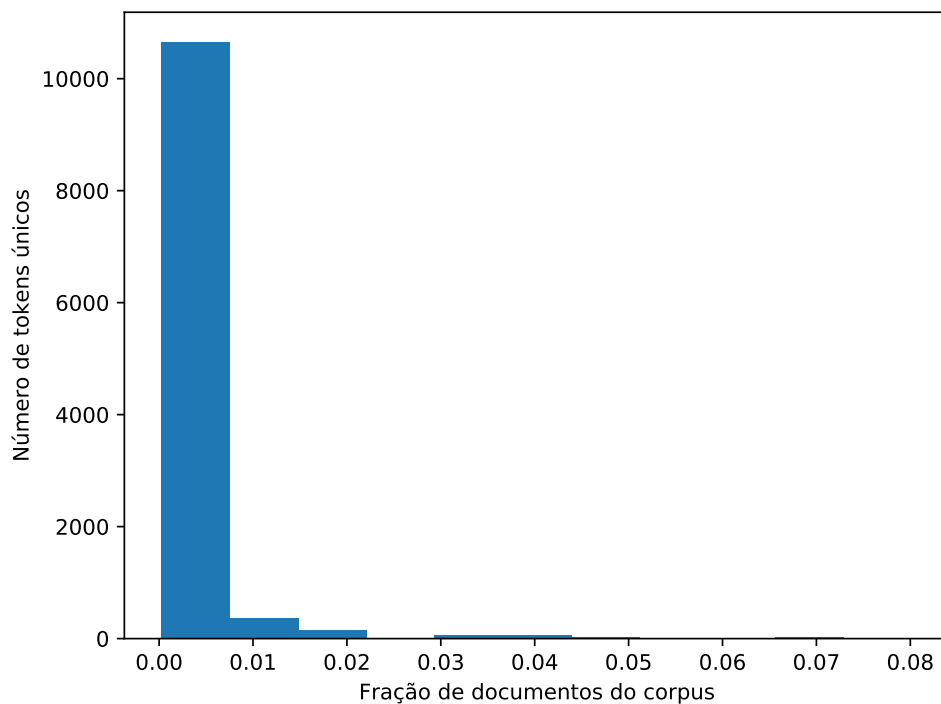


Figura 10 – Histograma de palavras da faixa de frequência de interesse a ser filtrada no *corpus* em português.

citar novamente o fato de que não existem *tokens* extremamente predominantes no *corpus*.

Tabela 3 – Faixas de frequência em documentos e número de *tokens* únicos existentes para o *corpus* em inglês.

Porcentagem dos documentos	Número de <i>tokens</i>
De 0% até 10%	29003
Acima de 10% até 20%	77
Acima de 20% até 30%	18
Acima de 30% até 40%	10
Acima de 40% até 50%	5
Acima de 50% até 60%	3
Acima de 60% até 70%	1
Acima de 70% até 80%	0
Acima de 80% até 90%	0
Acima de 90% até 100%	0

A Figura 11 exibe o histograma de análise para a faixa de frequência entre 0% e 10%. Novamente, nota-se que *tokens* com frequência inferior a 0,5% são os mais prevalentes. Assim como no caso anterior, esta faixa foi a escolhida para remoção. *Tokens* com alta frequência não foram removidos, já que sua presença não se dá neste *corpus*. Em razão da escolha de faixa de frequência, *tokens* que apareceram em menos de 160 documentos foram removidos. Finalmente, o vocabulário processado passou a contar com $|V| = 1524$ *tokens*.

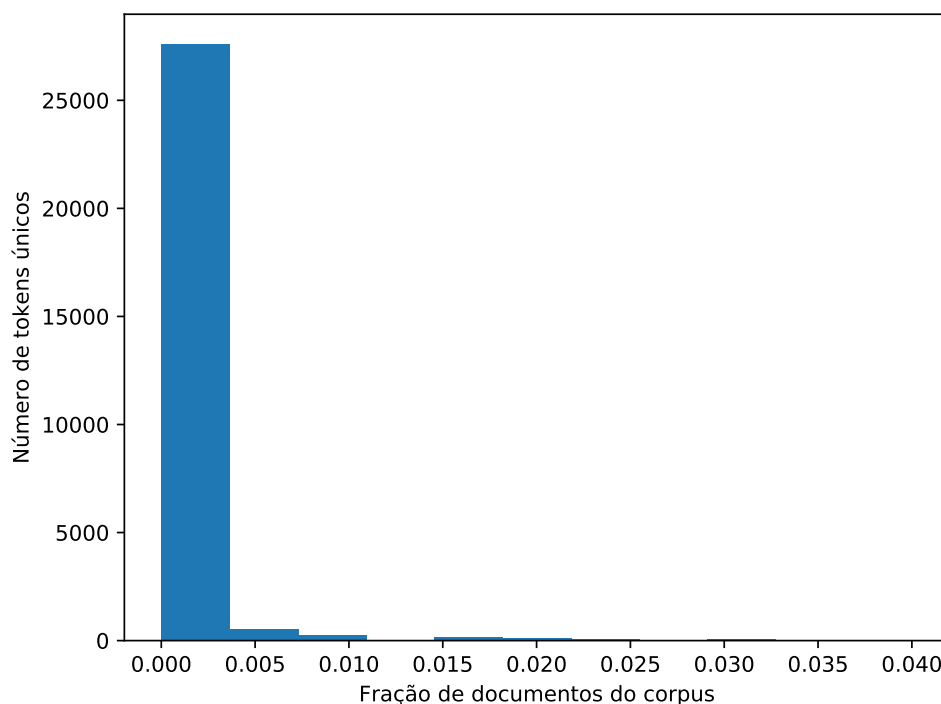


Figura 11 – Histograma de palavras da faixa de frequência de interesse a ser filtrada no *corpus* em inglês.

3.3 Preparação de recursos

Dado o *corpus* pré-processado segundo as etapas anteriores, foi necessária a realização de uma preparação dos recursos a serem usados no treinamento de cada um dos tipos de modelo: LDA, CTM e ETM. Isto foi necessário já que cada um dos modelos possui um formato de entrada apropriado e necessita de recursos distintos para funcionamento. Nesta etapa, os recursos compartilhados e exclusivos necessários ao treinamento de cada tipo de modelo foram produzidos. Os modelos compartilham entre si o vocabulário e dicionário de palavras, produzidos segundo as etapas descritas anteriormente. O dicionário é empregado no momento de contabilização dos valores de coerência de tópicos para cada um dos modelos após cada treinamento, o que é detalhado na subseção seguinte.

Além de vocabulário e dicionário, os modelos também precisam compartilhar o *corpora* para treinamento e validação. O *corpus* pré-processado foi dividido em duas partes, sendo essas um conjunto para treino dos modelos e outro para validação, destinado ao cálculo da medida de coerência. A métrica de coerência calculada serve de base para a escolha dos modelos a serem avaliados de forma mais aprofundada, para ambos os *corpora*. Assim como as demais etapas descritas ao longo deste trabalho, a divisão foi feita no *corpus* de língua portuguesa e no de língua inglesa. Em ambos, determinou-se que 80% dos documentos seriam destinados ao treinamento dos modelos, enquanto os 20% restantes seriam usados para validação. Essa abordagem foi implementada em razão

do maior interesse desse estudo pela avaliação subjetiva dos tópicos gerados. Caso uma abordagem como a validação cruzada fosse utilizada, seria confuso avaliar qualitativamente os resultados oriundos dos múltiplos conjuntos empregados. A Tabela 4 detalha o número de documentos não-vazios destinados para treino e validação em cada *corpus*. *Tokens* inexistentes no vocabulário foram removidos de ambos os conjuntos de documentos de cada idioma com o auxílio das instâncias de **Dictionary** do *gensim* formadas com o vocabulário produzido na seção acima. Documentos que tornaram-se vazios em razão da remoção de termos inexistentes no vocabulário foram removidos antes da divisão em conjuntos de treinamento e validação. A Tabela a seguir também indica o número de documentos vazios removidos em cada linguagem.

Tabela 4 – Divisão de documentos para treino e validação em ambos os *corpora* estudados.

<i>corpus</i>	documentos não vazios (pré-processados)	documentos vazios	treino	validação
português	3394	10	2715	679
inglês	32092	73	25674	6418

Em seguida, os recursos específicos de cada modelo foram produzidos. No caso do LDA, os recursos previamente criados já foram suficientes para seu treinamento. Portanto, esta etapa concentrou-se na preparação de recursos para os demais modelos. No caso do CTM, foi necessário produzir as entradas BOW e SBERT (REIMERS; GUREVYCH, 2020) para o modelo com base nos conjunto de dados separado para treinamento. Para criação do BOW, existe uma limitação padrão de que o vocabulário tenha no máximo 2000 *tokens*, isto é, $|V| = 2000$. No contexto deste estudo, a limitação não foi aplicada em nenhum dos dois vocabulários utilizados, já que os mesmos não excederam o tamanho citado. Para produzir a entrada SBERT, é empregado um conjunto de *embeddings* que pode ser escolhido mediante as opções disponíveis diretamente na *API* do CTM. Para o presente trabalho, para o *corpus* em português foi utilizado o modelo multilíngue de *embeddings* **distiluse-base-multilingual-cased**, indicado para modelagem de tópicos em idiomas distintos do inglês. Este modelo é pré-treinado em mais de 50 línguas, incluindo português. Por outro lado, para o *corpus* em inglês foi usado o modelo treinado exclusivamente em inglês chamado **bert-base-nli-mean-tokens**. Ambos os modelos utilizados são oriundos do pacote *sentence-transformers* (REIMERS; GUREVYCH, 2020; REIMERS; GUREVYCH, 2019).

Por outro lado, o modelo ETM necessita de *embeddings word2vec*, além da entrada BOW. Nesse caso, o modelo de *embeddings* do *Wikipedia2Vec* foi empregado. O *Wikipedia2Vec* é uma ferramenta para extração de *embeddings* de palavras da *Wikipedia*, que também tem a capacidade de extrair *embeddings* que representam entidades: conceitos com páginas próprias na enciclopédia livre (YAMADA et al., 2020). O *Wikipedia2Vec* fornece modelos *word2vec* pré-treinados no *corpus* da *Wikipedia* em inglês e em português,

e estes foram utilizados para entrada no modelo ETM. *Embeddings* com 300 dimensões foram empregados para os treinamentos de ambos os idiomas. Para o presente estudo, apenas os *embeddings* de palavras presentes no *Wikipedia2Vec* foram empregados, já que o conceito de entidade não está sendo trabalhado nesta monografia. Dessa forma, um pré-processamento foi realizado nos *embeddings* de forma a remover os vetores associados às entidades antes de seu uso pelo ETM. Além disso, o modelo ETM internamente considerava apenas os *embeddings* de termos existentes no vocabulário de treinamento recebido. Dessa forma, apenas o subconjunto dos *embeddings* referente ao vocabulário foi passado para o ETM. O carregamento dos *embeddings* em memória foi realizado com o uso da classe **KeyedVectors** do *gensim*, destinada para manipulação de vetores de palavras.

3.4 Treinamento dos modelos

Os recursos produzidos por meio da etapa anterior foram usados como entrada nos *scripts* de treinamento associados a cada modelo, produzindo e persistindo os resultados de treino. Nesse contexto, o primeiro fator a ser considerado no treinamento é o número K de tópicos, já que este é o principal hiperparâmetro de treinamento em modelagem de tópicos. Quando o *corpus* tem características conhecidas, é possível escolher uma faixa de valores para K que esteja relacionada à dimensionalidade temática real do conjunto de dados. Entretanto, no presente estudo, não existia conhecimento prévio sobre os *corpora* de forma que uma definição direta desses valores pudesse ser realizada. Dessa forma, uma faixa de valores para K foi definida para exploração nos modelos treinados. Em razão das diferenças arquiteturais entre os modelos LDA, CTM e ETM, apenas a variação do hiperparâmetro K foi realizada, e demais hiperparâmetros de cada modelo foram utilizados em suas configurações padrão, conforme fornecido nas referidas bibliotecas ou conforme descrito neste texto. Os valores para o hiperparâmetro K explorados neste estudo foram: $K \in \{5, 8, 10, 12, 15, 18, 20, 22, 25, 28, 30\}$. Um *script* de treinamento para cada um dos tipos de modelos foi criado, e cada treinamento foi realizado separadamente. Ao total, considerando todas as combinações de K e os três tipos de modelo, 33 modelos foram treinados usando o *corpus* para cada um dos idiomas, totalizando 66 modelos.

Os *scripts* de treinamento possuem codificação semelhante entre si, composta do carregamento de recursos necessários e do treinamento dos modelos propriamente ditos. Os modelos são treinados com o conjunto de dados de treino, e ao fim do processo, o valor de coerência de tópicos é calculado a partir dos dados de validação. A subseção a seguir detalha a métrica de coerência de tópicos utilizada. Além disso, ao fim de cada treinamento são obtidos os tópicos obtidos em forma de listas de palavras, onde cada lista representa um tópico. Os 20 termos mais relevantes de cada tópico são considerados nesta etapa. As probabilidades dos termos dentro de cada tópico também são extraídas neste momento. Um objeto representando os tópicos do modelo e as probabilidades de termos é

persistido para posterior análise, por meio da biblioteca *joblib* (VAROQUAUX; GRISEL, 2009). O nome do modelo, o valor de K utilizado em seu treino, o diretório onde sua representação foi salva e o valor de coerência de tópicos calculado para o mesmo são salvos como uma linha em um arquivo CSV⁶. Este arquivo contém as informações referentes à cada modelo de certo tipo treinado – LDA, CTM ou ETM. Os objetos persistidos e os arquivos CSV são utilizados na etapa de análise dos resultados.

O treinamento do LDA utilizou a implementação da classe **LatentDirichletAllocation**, existente no *scikit-learn* (PEDREGOSA et al., 2011). O LDA utilizou a inferência VB *online* para o aprendizado dos tópicos. No caso do CTM, a implementação do modelo existente na biblioteca *contextualized-topic-models*⁷ foi utilizada. Para este modelo, foi utilizada versão *CombinedTM*, indicada para modelagem de tópicos comum. Na API do modelo foi definido que a entrada SBERT do mesmo teria 512 dimensões para o treinamento usando o *corpus* em português e 768 dimensões para o *corpus* em inglês. As diferenças em tamanho de dimensões devem-se aos modelos de *embeddings* distintos usados em cada caso. Um total de 100 *epochs* – número de passagens do *corpus* pelo modelo durante treinamento – foi realizado, conforme a configuração padrão do modelo. Por fim, para o modelo ETM uma biblioteca foi produzida a partir do código original do modelo disponibilizado pelos seus autores. A biblioteca do ETM foi criada por razões de praticidade, já que o código original do modelo foi adaptado em muitos pontos para a realização do presente projeto ao longo do mesmo. A implementação da biblioteca ETM apenas adapta as funcionalidades básicas do modelo para serem acessíveis via API por meio de uma classe, e não faz mudanças significativas em sua arquitetura interna. Neste trabalho, a classe **ETM** da biblioteca foi utilizada como implementação do modelo. Em relação ao número de *epochs*, foi utilizado o número de 300 ao invés do padrão 20. Durante testes preliminares, o valor de 300 *epochs* trouxe resultados interpretáveis. A biblioteca ETM encontra-se disponível para acesso via Github⁸.

Métrica de coerência de tópicos

Diferentes abordagens para avaliação de qualidade de modelos de tópicos já foram propostas na literatura (LAU; NEWMAN; BALDWIN, 2014). Naturalmente, o “padrão-ouro” com o qual tais métodos são comparados é o julgamento humano. Estudos mais recentes indicam que a métrica de **coerência** possui alta correlação com os resultados provenientes de observações humanas (RÖDER; BOTH; HINNEBURG, 2015). A coerência de um tópico consiste de um cálculo indicativo do quanto os termos associados ao tópico ocorrem mutuamente, o que demonstra o nível de pertencimento de ambos a um mesmo tema. Logo, tópicos onde os termos ocorrem juntos com alta frequência possuem valores

⁶ *Comma-Separated Values*: arquivos textuais separados por vírgula, representando uma tabela de dados.

⁷ <<https://github.com/MilaNLPProc/contextualized-topic-models>>

⁸ <<https://github.com/lffloyd/embedded-topic-model>>

de coerência mais altos, e caso contrário, possuem coerência baixa. Dessa forma, pode-se considerar que a coerência indica o nível de interpretabilidade semântica associado ao tópico. Normalmente, a coerência é calculada com base nos termos mais relevantes para o tópico em questão, e leva em consideração o cálculo de uma métrica de confirmação para cada um dos pares de palavras do tópico. Por fim, os cálculos realizados para os pares de palavras são consolidados num valor final que ilustra a coerência do tópico. Quanto maior este valor, maior o nível de coerência associado. Existem diversas métricas de confirmação para cálculo de coerência, como a baseada em informação mútua pontual normalizada, do inglês *normalized pointwise mutual information* (NPMI) (RÖDER; BOTH; HINNEBURG, 2015). A métrica baseada em NPMI foi a escolhida para a avaliação quantitativa de qualidade dos modelos treinados neste trabalho.

O valor da coerência NPMI pode variar no intervalo $[-1, 1]$. Considerando alguns valores específicos, caso o valor obtido seja -1 , significa que não existe co-ocorrência entre o par de termos (a, b) avaliado. Por outro lado, se o valor obtido for 0 , significa que os termos co-ocorrem de forma aleatória. Por fim, se o valor calculado para a métrica for 1 , isso significa que os termos têm completa co-ocorrência (MIGDAL, 2015). Para realização do cálculo de coerência, (RÖDER; BOTH; HINNEBURG, 2015) propõem uma *pipeline* de tarefas com o propósito de formular como a métrica é calculada. Esta *pipeline* é formada por quatro etapas, a seguir: segmentação dos subconjuntos de palavras, estimação de probabilidades de ocorrência, cálculo da métrica de confirmação para um par de termos (a, b) e agregação das métricas de confirmação em um único valor. Para o presente trabalho, a implementação realizada pelo *gensim* da *pipeline* de coerência descrita foi empregada para o cálculo da métrica. A implementação da biblioteca é consolidada pela classe **CoherenceModel**, e possui diferentes formas de uso e possibilita o emprego de variadas métricas de confirmação além do NPMI (ŘEHŮŘEK; SOJKA, 2019). Para cálculo da coerência para os modelos treinados, foi necessário utilizar os tópicos gerados pelo modelo, o conjunto de documentos de validação criado especificamente para este passo, o dicionário de palavras do vocabulário e também o número de palavras a serem consideradas como mais relevantes dentro do tópico. Nesse caso, foi usado o mesmo número de palavras retornadas por tópico ao treinar cada modelo: 20 palavras.

3.5 Análise de resultados

Na maioria das etapas do estudo descritas até o momento, não fez-se distinção relevante entre o tratamento dos conjuntos de dados em português e em inglês, salvo algumas exceções. Contudo, a análise dos resultados obtidos deve considerar as características de cada *corpus*, que vão além da diferença linguística entre eles. Deve-se considerar também a disparidade do tamanho de cada *corpus*, assim como as diferenças em constituição interna, referentes ao processo de coleta de dados diferente entre ambos. Portanto, uma análise

dos resultados do presente estudo deve considerar múltiplos aspectos com o objetivo de consolidar os achados apresentados da forma mais completa possível.

De forma a identificar os modelos com maior coerência associada para um aprofundamento de sua análise, a métrica de coerência servirá como base de filtragem. Para cada modelo, serão considerados os valores calculados para coerência de tópicos usando o conjunto de documentos destinado a validações, com o intuito de determinar aqueles modelos que aprenderam tópicos mais interpretáveis. Para ambos os idiomas, o modelo com melhor valor para a métrica de coerência será escolhido para a continuidade de sua análise. A análise dos modelos assim definidos será realizada em dois âmbitos distintos, descritos a seguir:

- **análise qualitativa** – os modelos com melhores resultados de coerência serão avaliados qualitativamente no contexto dos tópicos gerados. A capacidade de rotulação dos tópicos nesta etapa será levada em consideração, de forma a investigar empiricamente se os modelos geraram tópicos interpretáveis;
- **análise léxica** – os tópicos dos modelos escolhidos serão avaliados quanto às categorias lexicais associadas aos mesmos. Dessa forma, as ferramentas LIWC e *Empath* serão utilizadas. Versões mais recentes do LIWC, como a de 2015, não permitem o uso de seus dicionários internos em conjunto com bibliotecas externas para manipulação de seu conteúdo. Como os dicionários em inglês do *software* não podem ser acessados, o LIWC foi utilizado apenas para a análise léxica dos tópicos obtidos pelo modelo treinado no *corpus* de língua portuguesa do *Reddit*. O dicionário LIWC 2007 em português é acessível no portal do LIWC, e seu uso pode ser feito de forma facilitada com a biblioteca *liwc-python* (BROWN, 2012). Por outro lado, versões mais recentes do dicionário em português, apesar de disponíveis para uso, são incompatíveis com a ferramenta citada. Dessa forma, a versão de 2007 do dicionário foi a utilizada. Por sua vez, o *Empath* é treinado apenas em *corpus* de língua inglesa, o que também contribuiu para sua escolha como ferramenta de análise léxica para os tópicos extraídos pelo modelo treinado no *corpus* em inglês da rede social. Neste contexto, a biblioteca Python *empath-client* foi empregada para realização da análise (FAST, 2016). Em ambos os casos, uma discussão sobre as categorias existentes nos conjuntos de tópicos será feita, de forma a aprofundar as constatações propostas durante a análise qualitativa.

O capítulo seguinte irá explorar as dimensões citadas para avaliação dos resultados obtidos, destacando o que foi percebido como características dos modelos treinados em ambos os idiomas. Além da análise individual dos achados obtidos em cada linguagem, as semelhanças e diferenças entre os resultados dos treinamentos em português e em inglês serão exploradas.

4 Resultados

Neste capítulo, os resultados obtidos com a metodologia proposta no capítulo anterior são explorados. Uma seção foi destinada para cada língua com o objetivo de pormenorizar os resultados de treinamento observados. Por meio da métrica de coerência de tópicos, foi possível identificar os modelos treinados com maior coerência associada, em ambos os idiomas. Os valores da métrica de coerência foram calculados para cada modelo após seu treinamento, com base nos documentos separados para validação. Como descrito anteriormente, para cada valor de número de tópicos (K) definido, um modelo de cada uma das três arquiteturas foi treinado, em ambos os *corpora*. Considerando que $K \in \{5, 8, 10, 12, 15, 18, 20, 22, 25, 28, 30\}$, existem 11 valores possíveis para o hiperparâmetro. Portanto, 33 modelos diferentes foram treinados em cada idioma, totalizando 66 modelos treinados. Os modelos com maior coerência, em cada linguagem, foram escolhidos para um aprofundamento de seu estudo: uma análise em duas camadas foi realizada para proporcionar maior riqueza de interpretação dos resultados obtidos nesses modelos. Contudo, a escolha dos modelos com base nos valores de coerência não tem como objetivo determinar que um modelo é melhor que os demais, e sim de proporcionar um direcionamento para o escopo de análise subjetiva. Sendo assim, primeiramente, uma **análise qualitativa** teve como objetivo identificar subjetivamente a semântica associada aos tópicos extraídos pelos modelos definidos na etapa anterior. Em seguida, uma **análise léxica** dos tópicos foi empregada de forma complementar, utilizando para tal as ferramentas de análise textual LIWC e *Empath*, respectivamente para os tópicos em português e em inglês. As seções a seguir detalham os achados levantados por meio das etapas citadas para cada um dos *corpora* de treinamento explorados. As visualizações de nuvens de palavras apresentadas neste capítulo levam em consideração as probabilidades de cada termo em cada tópico para determinar a proporção das representações. Sendo assim, palavras com maior probabilidade dentro de um tópico serão exibidas em tamanho maior nas nuvens de palavras.

4.1 Resultados no *corpus* em português

4.1.1 Desempenho dos modelos em português segundo a coerência

Os modelos de tópicos treinados no *corpus* em língua portuguesa foram listados em ordem decrescente de coerência calculada com o conjunto de dados de validação, considerando a variação apenas na quantidade de tópicos (K). As Tabelas 5, 6 e 7 relacionam os cinco modelos com maiores valores de coerência para as arquiteturas LDA, CTM e ETM, respectivamente. Além dos valores de coerência, as tabelas também listam

os tempos de treinamento referentes ao treinamento de cada modelo, em segundos. Os tempos mostrados são referentes à duração de execução dos treinamentos do ponto de vista de tempo de relógio, e não representam o tempo de processamento da CPU. Por sua vez, a Tabela 8 lista em ordem decrescente de valor de coerência os cinco modelos com melhor desempenho, dentre todos os 33 treinados com o *corpus* de língua portuguesa do *Reddit*.

Tabela 5 – Modelos LDA em português por ordem decrescente de coerência.

modelo	K	NPMI	tempo (s)
<i>lda_ptk5</i>	5	-0,006473	6,027192
<i>lda_ptk8</i>	8	-0,017459	5,429506
<i>lda_ptk10</i>	10	-0,047824	5,168725
<i>lda_ptk12</i>	12	-0,059560	5,302516
<i>lda_ptk15</i>	15	-0,073969	5,485522

Tabela 7 – Modelos ETM em português por ordem decrescente de coerência.

modelo	K	NPMI	tempo (s)
<i>etm_ptk5</i>	5	-0,025955	66,121439
<i>etm_ptk8</i>	8	-0,035100	66,568795
<i>etm_ptk10</i>	10	-0,039160	66,166846
<i>etm_ptk12</i>	12	-0,043647	67,162926
<i>etm_ptk15</i>	15	-0,055317	66,922859

Tabela 6 – Modelos CTM em português por ordem decrescente de coerência.

modelo	K	NPMI	tempo (s)
<i>ctm_ptk20</i>	20	-0,188456	75,464190
<i>ctm_ptk28</i>	28	-0,203868	76,100783
<i>ctm_ptk10</i>	10	-0,206100	74,578925
<i>ctm_ptk25</i>	25	-0,211983	75,876350
<i>ctm_ptk15</i>	15	-0,218820	75,388322

Tabela 8 – Modelos em português por ordem decrescente de coerência.

modelo	K	NPMI	tempo (s)
<i>lda_ptk5</i>	5	-0,006473	6,027192
<i>lda_ptk8</i>	8	-0,017459	5,429506
<i>etm_ptk5</i>	5	-0,025955	66,121439
<i>etm_ptk8</i>	8	-0,035100	66,568795
<i>etm_ptk10</i>	10	-0,039160	66,166846

Inicialmente, pode-se notar nas Tabelas 5, 6 e 7 que os valores de coerência para os modelos treinados situaram-se em sua totalidade no intervalo $[-1, 0)$. Segundo a definição da métrica de coerência NPMI, um valor próximo de 0 indica co-ocorrência aleatória entre os termos de cada tópico. No conjunto de modelos acima, os valores de coerência situaram-se predominantemente abaixo e próximos de 0, o que pode indicar uma leve tendência à baixa co-ocorrência dos termos associados a cada tópico. Dado o contexto do treinamento usando uma base de dados de postagens em português relativamente pequena, os valores de coerência mostram uma tendência maior à fragmentação de assuntos presentes no *corpus*. Deve-se notar que o *corpus* utilizado para composição da base de dados em língua portuguesa não foi originado utilizando um *subreddit* específico ao tema depressão na rede social, e sim um *subreddit* mais generalista. Portanto, os resultados condizem com o que espera-se do *corpus* utilizado. Outro ponto observável é a diferença acentuada nos tempos de treinamento calculados para cada tipo de modelo, mesmo com um conjunto de dados reduzido. Enquanto os modelos LDA listados foram treinados com tempo médio aproximado de $t \approx 5s$, os modelos CTM e ETM levaram muito mais tempo: respectivamente, $t \approx 75s$ e $t \approx 66s$. As diferenças de tempo podem ser atribuídas à maior complexidade arquitetural dos modelos CTM e ETM, que envolvem o emprego de modelos de representação de linguagem usando *embeddings*.

Considerando a Tabela 8, nota-se sobretudo a dominância dos modelos ETM nas primeiras posições, apesar de o modelo LDA com 5 tópicos ter sido aquele com maior

coerência. Ainda, a Figura 12 exibe a evolução da média de coerência para cada um dos tipos de modelos de tópicos à medida que o valor de K é incrementado. Nota-se que considerando os modelos LDA e ETM, os valores de coerência tiveram decréscimo com o aumento do número de tópicos. Isso indica que a diversidade de tópicos distintos no *corpus* explorado é baixa, já que quanto menor o número de tópicos, maior a coerência associada. Pode-se notar também que os modelos CTM tiveram valores de coerência consistentemente mais baixos que os apresentados pelos demais modelos. Os *embeddings* usados para treinamento dos modelos CTM não foram treinados com o idioma português exclusivamente, sendo constituídos por um aprendizado multilíngue, o que pode ajudar a explicar os resultados apresentados por este modelo. Além disso, percebe-se que com exceção dos casos onde $K < 10$, os modelos ETM apresentaram resultados melhores em termos de coerência em comparação aos demais. Isto indica o ganho semântico obtido com o uso de representações vetoriais da linguagem como o *word2vec*: mesmo com um *corpus* caracterizado por poucos tópicos semanticamente distintos entre si, o ETM consegue manter níveis de coerência superiores ao LDA ao incrementar-se o valor de K .

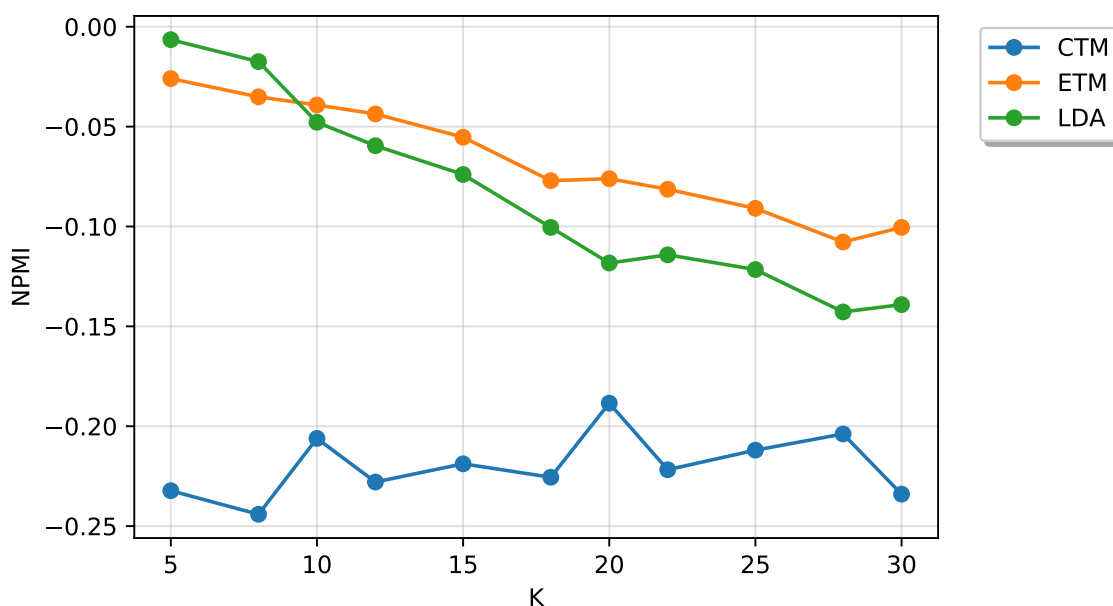


Figura 12 – Variação da coerência NPMI à medida que o número de tópicos é incrementado nos modelos em português.

Dentro do contexto do *corpus* de língua portuguesa, a instância *lda_ptk5* foi aquela que obteve melhor valor para a métrica de coerência. Desta forma, essa será a instância a ser analisada nas fases posteriores da presente análise.

4.1.2 Análise qualitativa

A Tabela 9 a seguir exibe os tópicos extraídos pelo modelo LDA treinado no *corpus* em português com $K = 5$, em ordem decrescente de coerência. Note que apenas os dez

primeiros termos de cada tópico são listados na tabela, de forma a reduzir o tamanho da representação. Além disso, percebe-se que certos tópicos possuem difícil interpretação semântica de seu significado. Por exemplo, o tópico T_4 reúne termos como “suicídio” e “problema”, mas o restante das palavras pertencentes ao mesmo não evocam semântica semelhante a estes dois termos. O valor para a métrica de coerência deste tópico corrobora a observação, já que é o menor valor dentre os tópicos do modelo. Comumente, tópicos desse tipo aparecem na modelagem de tópicos. Tópicos que agregam termos com pouca relação entre si são chamados de *junk topics* (ALSUMAIT et al., 2009). Por outro lado, tópicos com maior consistência semântica também são notados no modelo avaliado. As Figuras 13, 14, 15, 16 e 17 exibem as nuvens de palavras dos tópicos extraídos pelo modelo. As nuvens de palavras são compostas pelos vinte termos mais relevantes dentro de cada tópico, definidos pelas probabilidades associadas a cada uma das palavras dentro de suas distribuições de tópicos.

Tabela 9 – Tópicos do modelo LDA com $K = 5$ em ordem decrescente de coerência.

tópico	palavras	coerência
T_0	pai casar vidar ano coisa mae pessoa familia dia empregar	-0,002456
T_2	vidar dia ansiedade coisa ano tempo problema tratamento ajudar pessoa	-0,004057
T_3	ano dia coisa amigo tempo vidar pessoa gente casar namorar	-0,004769
T_1	ano pai faculdade dinheiro casar cursar dia trabalhar empresar cidade	-0,005065
T_4	pessoa vidar coisa tempo mundo suicidio problema formar vezar dia	-0,016018

Uma observação mais minuciosa dos tópicos depreende informações relevantes para o estudo.

- T_0 : 7,42% **pai** + 5,05% **casar** + 3,6% **vidar** + 3,47% **ano** + 2,67% **coisa** ... – este tópico possui o maior valor de coerência e agrega termos relacionados especialmente ao conceito de família. Na Figura 14, palavras como “**pai**”, “**mae**” e “**familia**” são alguns dos termos que aparecem em maior tamanho, o que indica sua maior contribuição para o tema latente avaliado. Discutivelmente, a qualidade das relações familiares de um indivíduo pode impactar o seu bem-estar psicológico. Em (WANG et al., 2020), os autores mostram que disfunção familiar está associada a níveis maiores de depressão e ansiedade em adolescentes. A dificuldade em compartilhar emoções dentro da família caracteriza esse tipo de lar, o que, de acordo com o trabalho citado, prejudica a capacidade dos jovens em obter suporte em momentos difíceis, agravando as chances de desenvolvimento de depressão e ansiedade. Por outro lado, a aparição deste tópico pode indicar uma busca por apoio familiar por parte do indivíduo. Dessa forma, a temática pode estar relacionada ao desejo de aproximação familiar, para compartilhamento de angústias e troca de experiências, ou mesmo a relatos sobre experiências emocionalmente positivas de acolhimento;
- T_1 : 2,83% **ano** + 2,13% **pai** + 1,82% **faculdade** + 1,67% **dinheiro** + 1,66% **casar** ... – este tópico associa palavras que podem ser divididas em ao menos dois

Figura 13 – T_0 do modelo LDAFigura 14 – T_1 do modelo LDAFigura 15 – T_2 do modelo LDAFigura 16 – T_3 do modelo LDAFigura 17 – T_4 do modelo LDA

temas relacionados: vida acadêmica e situação financeira. A presença de palavras como “**faculdade**”, “**cursar**” e “**aula**”, entre outras, indicam o primeiro tema, enquanto termos como “**dinheiro**” e “**trabalhar**” indicam o segundo. Estudos diversos mostram que a prevalência de depressão entre estudantes universitários é alta (SAROKHANI et al., 2013; BEITER et al., 2015; FERNANDES et al., 2018). (BEITER et al., 2015) mostram que o desempenho e sucesso acadêmicos, além da situação financeira, são alguns dos fatores que causam maior preocupação nos estudantes. Apesar da temática discernível, o tópico é aquele com segundo menor valor para coerência, o que é reforçado pela existência de termos com pouca conexão com o restante das palavras do tópico, como “**partir**”, “**dia**” e “**mundo**”;

- T_2 : 2,36% **vidar** + 2,25% **dia** + 2,24% **ansiedade** + 1,85% **coisa** + 1,81%

ano ... – tópico com segundo maior valor de coerência, associando palavras que evocam tratamento de doença mental. Termos como “**ansiedade**”, “**problema**”, “**tratamento**” e “**remédios**” são de grande relevância para o tópico. O uso de medicações controladas e o acompanhamento médico constante podem ser fatores que desestimulam a busca e o tratamento adequado da depressão. (SERNA et al., 2010) mostraram que grande parte dos pacientes abandona o tratamento logo nos primeiros meses de acompanhamento, e a identificação de tais casos pode ajudar a evitar insucessos no tratamento;

- T_3 : 5,02% **ano** + 3,63% **dia** + 3,06% **coisa** + 3,05% **amigo** + 2,61% **tempo** ... – tópico com valor de coerência intermediário, que reúne termos associados a relacionamentos interpessoais: a existência de palavras como “**amigo**”, “**namorar**”, “**pessoa**” e “**gente**” trazem tal interpretação. O surgimento deste tema é relevante para o estudo, já que comumente indivíduos depressivos encontram maiores dificuldades em se relacionar com outras pessoas (TRISCOLI; CROY; SAILER, 2019). Além disso, o tópico pode representar a busca por relacionamentos emocionalmente significativos, com o intuito de proporcionar ao indivíduo apoio e suporte emocional. Ainda assim, existem no tópico outros termos com pouco significado compartilhado e em tamanhos maiores, indicando maior importância para o tópico, como “**dia**”, “**coisa**” e “**ano**”;
- T_4 : 7,49% **pessoa** + 4,33% **vidar** + 3,67% **coisa** + 1,63% **tempo** + 1,47% **mundo** ... – tópico com o menor valor de coerência dentre os extraídos pelo modelo. Termos com pouca ligação clara entre si são agregados neste tópico, como os pares “**mundo**” e “**atar**”, ou “**formar**” e “**problema**”. Contudo, mesmo nesse tópico surgem palavras relevantes para o tema da depressão: “**medo**”, “**sentir**” e “**suicídio**”.

4.1.3 Análise léxica

O modelo LDA com $K = 5$ também teve seus tópicos avaliados quanto às categorias lexicais de seus termos. A biblioteca *liwc-python* foi utilizada em conjunto com o dicionário de 2007 do LIWC em português para determinação das categorias de palavras dos tópicos (FILHO; PARDO; ALUÍSIO, 2013). Além disso, o manual de propriedades psicométricas do LIWC foi empregado para identificação das categorias a partir de suas abreviações (PENNEBAKER et al., 2007). De forma a ter uma visão ampla sobre as categorias de palavras agregadas no modelo, uma visualização das categorias foi criada. A visualização descreve as frações do conjunto de palavras dos tópicos correspondentes às categorias-padrão existentes no LIWC. A visualização é exibida na Figura 18. Apenas as cinco categorias mais predominantes nos tópicos são exibidas, e apenas os vinte termos mais relevantes dentro de cada tópico foram considerados nesta agregação. É importante notar que uma mesma palavra no LIWC pode ser incluída em diversas categorias lexicais.

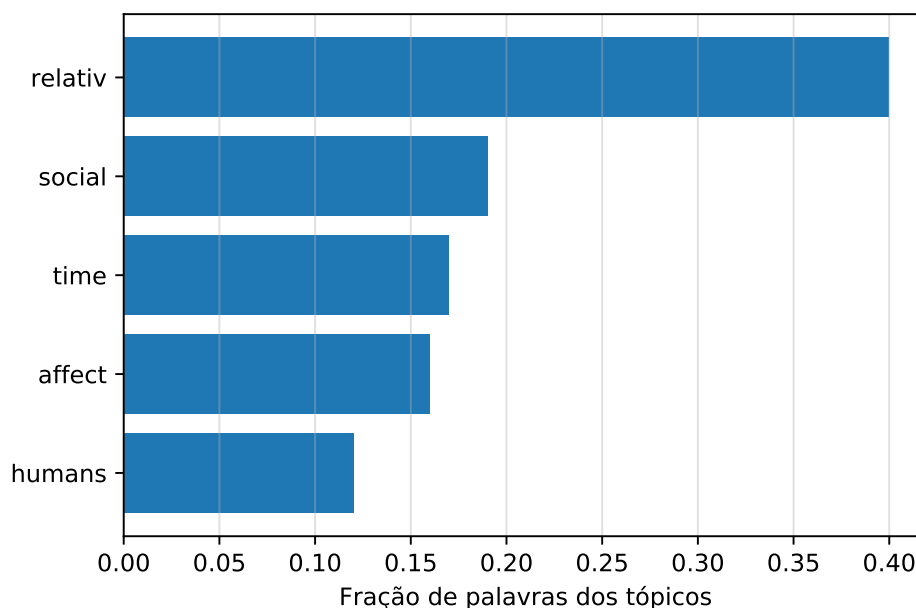


Figura 18 – Cinco categorias lexicais predominantes nos tópicos do modelo LDA com $K = 5$ treinado em português, segundo o LIWC.

Ao todo, o LIWC identificou a presença de 36 categorias lexicais nos tópicos avaliados. Percebe-se que a categoria predominante nos tópicos é *relativ*. Essa categoria corresponde ao conceito de relatividade no LIWC, responsável por agregar palavras associadas a orientação espacial, como “área”, “virar” ou “sair”. No caso dos tópicos analisados, o LIWC contabilizou palavras díspares como “trabalhar”, “cidade” e “ano”, entre outras, como pertencentes a esta categoria. Dessa forma, esse resultado é de difícil interpretação. Por outro lado, vê-se na Figura que a categoria *social* é a segunda mais predominante. Essa categoria está associada a processos sociais, como interação interpessoal por exemplo. Como termos associados a esta categoria, são exemplos as palavras “pessoa”, “filho”, “pai”, “ajudar”, entre outras. Nesse caso, a relação semântica entre a categoria descrita e os termos associados é perceptível. Palavras associadas ao conceito de família aparecem nesta categoria, o que tem associação com o tópico T_0 do modelo, que é representativo deste tema. A próxima categoria é *time*, associada a conceitos de passagem de tempo, e engloba palavras como “ano”, “dia” ou “mês”. Em todos os tópicos avaliados, palavras associadas ao conceito surgiram dentre as mais relevantes. Isto reforça que a temporalidade é um conceito frequente no *corpus* explorado. Ainda, a categoria seguinte é *affect*, que descreve processos afetivos diversos, envolvendo emoções de valoração positiva ou negativa. Nesta categoria, o LIWC associou palavras como “amigo”, “medo”, “problema” e “ansiedade”. No LIWC, existem especializações da categoria *affect* associadas respectivamente à processos de valoração emocional negativa e positiva, chamados de *negemo* e *posemo*. Contudo, os termos citados não foram categorizados de forma mais específica. De qualquer forma, pode-se notar a ligação entre as palavras e o conceito definido pela categoria. Além disso,

é notável a presença da palavra “problema” aqui, termo este que aparece em quatro dos cinco tópicos entre as palavras mais relevantes. Por fim, outra categoria predominante é a descrita por *humans*, que descreve conceitos intrinsecamente associados aos seres humanos. Para o conjunto de tópicos, esta categoria uniu palavras como “pessoa”, “filho” e “gente”. Nota-se grande semelhança entre os termos associados por esta categoria com os descritos na categoria *social*. A palavra “pessoa”, principalmente, aparece na maioria dos tópicos deste modelo.

4.2 Resultados no *corpus* em inglês

4.2.1 Desempenho dos modelos em português segundo a coerência

De maneira semelhante ao que foi feito com os modelos treinados com o *corpus* em português, os modelos do *corpus* em inglês também foram ordenados de acordo com os resultados de coerência obtidos para o conjunto de dados de validação. As Tabelas 10, 11 e 12 a seguir ilustram os resultados para os cinco modelos melhor colocados de cada arquitetura. Note que nas tabelas o tempo de treinamento é dado em minutos. Novamente, os tempos referem-se à duração de execução dos treinamentos do ponto de vista de tempo de relógio, e portanto não refletem o tempo de processamento da CPU. Ainda, a Tabela 13 ordena por forma decrescente de coerência os cinco modelos com melhor desempenho, dentro todos aqueles treinados com o *corpus* de língua inglesa do *Reddit*.

Tabela 10 – Modelos LDA em inglês por ordem decrescente de coerência.

modelo	K	NPMI	tempo (m)
<i>lda_enk15</i>	15	0,016794	1,06914
<i>lda_enk10</i>	10	0,009414	1,00168
<i>lda_enk12</i>	12	0,008639	1,02013
<i>lda_enk18</i>	18	0,003335	1,070438
<i>lda_enk8</i>	8	0,003221	0,945168

Tabela 12 – Modelos ETM em inglês por ordem decrescente de coerência.

modelo	K	NPMI	tempo (m)
<i>etm_enk28</i>	28	0,024751	50,560078
<i>etm_enk20</i>	20	0,022152	41,7813
<i>etm_enk25</i>	25	0,021671	42,778675
<i>etm_enk22</i>	22	0,021667	43,147416
<i>etm_enk30</i>	30	0,020555	56,015518

Tabela 11 – Modelos CTM em inglês por ordem decrescente de coerência.

modelo	K	NPMI	tempo (m)
<i>ctm_enk8</i>	8	-0.029757	24,724499
<i>ctm_enk30</i>	30	-0.035859	24,397375
<i>ctm_enk20</i>	20	-0.037332	23,839186
<i>ctm_enk28</i>	28	-0.038115	26,47678
<i>ctm_enk22</i>	22	-0.039757	24,712849

Tabela 13 – Modelos em inglês por ordem decrescente de coerência.

modelo	K	NPMI	tempo (m)
<i>etm_enk28</i>	28	0,024751	50,560078
<i>etm_enk20</i>	20	0,022152	41,7813
<i>etm_enk25</i>	25	0,021671	42,778675
<i>etm_enk22</i>	22	0,021667	43,147416
<i>etm_enk30</i>	30	0,020555	56,015518

Uma fator visível nos resultados neste *corpus* diz respeito aos valores de coerência obtidos. Majoritariamente, os valores de coerência descritos nas tabelas encontram-se no intervalo $(0, 1]$. A exceção foram os modelos CTM, cujos valores de coerência ficaram abaixo de 0. Na métrica NPMI, valores próximos de 0 indicam que a co-ocorrência de

termos é aleatória, mas a tendência levemente positiva – indicando levemente uma maior propensão à co-ocorrência entre termos dos tópicos – dos valores observados condiz com as características do *corpus* explorado. A base de dados foi construída mediante a coleta de postagens existentes no *subreddit depression*, destinado a conteúdo com foco único no tema de mesmo nome, o que naturalmente garante maior unidade temática às postagens reunidas. Além disso, o tamanho do *corpus* contribui para uma extração de tópicos de maior qualidade, já que a existência de um conjunto maior de dados estimula a capacidade de percepção de padrões dos modelos de tópicos. Adicionalmente, as diferenças em tempo de treinamento entre as arquiteturas apresentaram um incremento significativo. Os modelos treinados com o *corpus* em português eram treinados com tempos que variavam de segundos a alguns minutos. Aqui, os tempos ultrapassaram a marca de um minuto na grande maioria dos casos, atingindo aproximadamente uma hora em certos casos. Enquanto os modelos LDA foram treinados em aproximadamente $t \approx 1min$, os modelos CTM tiveram tempos por volta dos $t \approx 25min$. Por sua vez, os modelos ETM tiveram os treinamentos mais longos: no caso dos modelos listados, a média ficou próxima de $t \approx 47min$. Novamente, ficam claros os impactos das diferenças arquiteturais nos tempos de treinamento observados, o que foi acentuado pelo emprego de um *corpus* cerca de dez vezes maior que o empregado no treinamento de modelos em português.

O predomínio dos modelos ETM na Tabela 13 é notável, e a Figura 19 a seguir ajuda a explicar esses resultados. Para valores $K \leq 15$, os modelos LDA e ETM tiveram níveis bastante semelhantes de coerência. Contudo, para $K > 15$ os modelos LDA sofreram um decréscimo constante de coerência, enquanto os modelos ETM mantiveram o padrão de subida demonstrado até este momento por ambos os modelos. O gráfico possibilita constatar dois pontos: 1) os modelos ETM depreenderam maior semântica do *corpus* que os modelos LDA à medida que K foi incrementado, já que seus valores para coerência continuaram a subir; e 2) a elevação da coerência à medida que K cresce indica que o *corpus* trabalhado possui maior diversidade de tópicos latentes, se comparado à sua contraparte em português. A capacidade dos modelos ETM já havia sido notada na análise dos resultados baseados no *corpus* de língua portuguesa, apesar do mesmo não ter tido o melhor valor de coerência naquele caso. Contudo, este fator torna-se mais perceptível no caso do *corpus* em inglês. A observação sobre a diversidade de tópicos no *corpus* também possui consistência com as características do conjunto de dados: a maior quantidade de postagens agregadas no *subreddit* amplia os horizontes temáticos abarcados pelo *corpus*.

Neste contexto, o modelo *etm_enk28* foi a instância com maior coerência dentre aquelas treinadas e servirá como ponto central para a análise qualitativa que segue.

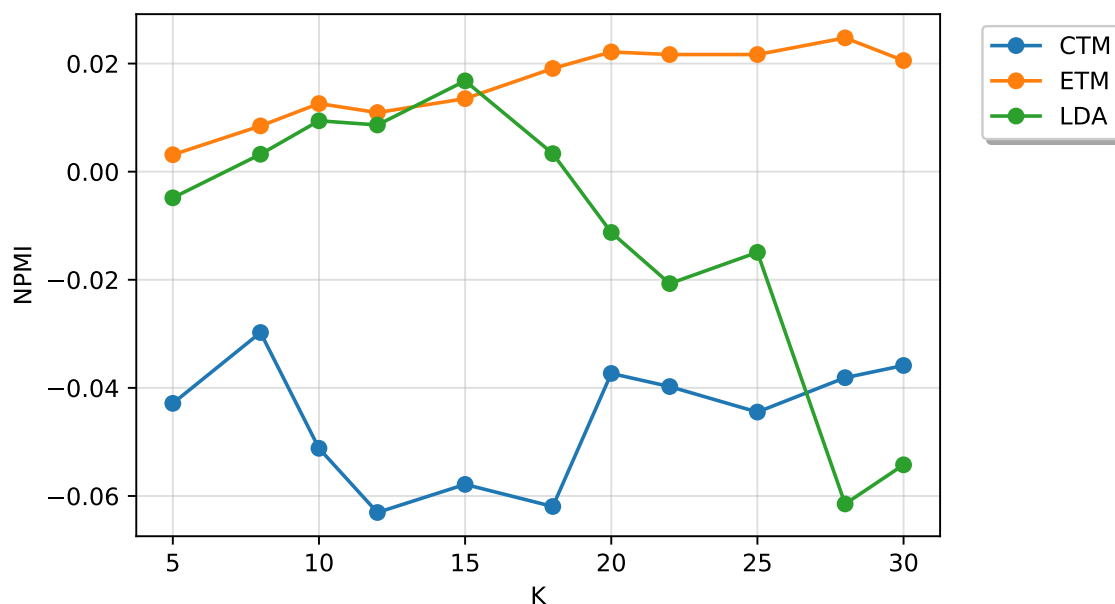


Figura 19 – Variação da coerência NPMI à medida que o número de tópicos é incrementado nos modelos em inglês.

4.2.2 Análise qualitativa

Dado o amplo número de tópicos do modelo ETM a ser avaliado de forma mais profunda, a análise qualitativa de seus tópicos será realizada em partes. Abaixo, a Tabela 14 exibe os valores calculados para coerência para cada tópico. Assim como nos demais casos deste capítulo, a tabela encontra-se ordenada de maneira decrescente a partir dos valores de coerência. Inicialmente, os resultados descritos na tabela contrastam-se de forma significativa com o observado no caso do modelo LDA analisado para o *corpus* em português. Enquanto naquele modelo os valores de coerência por tópico sempre ficavam abaixo do zero, no presente modelo a grande maioria dos tópicos teve coerência com valoração positiva.

Em razão da ampla quantidade de tópicos a serem avaliados, uma subseção foi criada para a avaliação de grupos de tópicos contendo oito elementos cada. Dessa forma, as respectivas figuras ilustrando cada tópico ficarão agrupadas na subseção onde o tópico é abordado especificamente. A última subseção contém apenas os quatro tópicos restantes para a análise neste modelo.

4.2.2.1 Análise qualitativa dos tópicos 0 a 7

Abaixo, os tópicos T_0 a T_7 são analisados em maiores detalhes.

- T_0 : 7, 34% **depression** + 4, 69% **anxiety** + 3, 2% **doctor** + 3, 11% **medication** + 2, 75% **mental** ... – primariamente, as palavras deste tópico são associadas ao tema de tratamento psicológico ou psiquiátrico. Na Figura 20 pode-se atestar isto, já que

Tabela 14 – Tópicos do modelo ETM com $K = 28$ em ordem decrescente de coerência.

tópico	palavras	coerência
T_5	school college year go class high parent grade fail study	0,096122
T_0	depression anxiety doctor medication mental med therapist therapy issue experience	0,090807
T_{22}	parent family year mom mother old dad get kid brother	0,071539
T_1	post read write reddit may know see story amp watch	0,060227
T_{23}	day sleep go night bed eat wake hour drink get	0,059759
T_{24}	say tell ask talk call know text question answer get	0,038913
T_{15}	life live die dream want hope world death happy future	0,038258
T_{21}	look see head cry walk face sit eye hand come	0,034008
T_{27}	hate fuck shit fucking want know stupid suck shitty end	0,027033
T_7	time use thing play make game enjoy try music new	0,026105
T_{19}	feel happy sad feeling cry know depressed bad emotion sadness	0,023777
T_{12}	year month last time past week start couple day new	0,02308
T_3	self pain fear life weight body experience may lose low	0,022105
T_4	love girl relationship guy want girlfriend break know meet find	0,019926
T_9	help depression need problem depressed struggle try stress find suffer	0,019301
T_{14}	work job money pay get life time year month need	0,017678
T_{25}	friend good talk go time make hang close get group	0,014195
T_{18}	talk feel think know good seem depressed like thank say	0,009897
T_{16}	would think could know wish thought thing well happen day	0,002019
T_{26}	work thing time try find feel seem thought hard motivation	0,001879
T_{10}	time take go day could first come end start would	0,001572
T_6	make know feel way try think reason end idea seem	0,000799
T_{17}	life one world way people person see make good thing	0,000131
T_{11}	people care person social know make think feel many problem	0,000059
T_{13}	want know go feel see need make tell let thing	0,000001
T_2	try suicide go keep stop fight kill care leave lose	-0,000063
T_8	go leave move live place find give come make see	-0,000451
T_{20}	get go bad start thing feel make think lot deal	-0,005651

palavras como “**depression**”, “**anxiety**”, “**doctor**” ou “**medication**” aparecem em maiores proporções em relação às demais, o que denota sua importância. O tópico tem especial importância já que, como dito anteriormente, a identificação de indivíduos desestimulados com a intervenção médica é extremamente importante para evitar o abandono do tratamento. Deve-se dizer ainda que este tópico apresenta o segundo maior valor de coerência, o que é atestado pela sua interpretabilidade;

- T_1 : 6,98% **post** + 4,58% **read** + 3,47% **write** + 2,37% **reddit** + 2,23% **may** ... – a Figura 21 ilustra esse tópico, que reúne palavras associadas à própria atividade de utilizar o *Reddit* em si. Sua aparição aqui tem um contraste interessante com os resultados observados no *corpus* em português, já que naquele caso um tópico semelhante a este não foi observado. O tópico representa uma ideia metalinguística, já que enquanto usuários utilizam a rede social para expor suas emoções e pensamentos, eles também abordam as mecânicas associadas ao próprio uso da rede. Termos como “**post**”, “**read**”, “**write**”, entre outros, ilustram isso. Nota-se que este tópico possui o quarto maior valor de coerência, o que é perceptível ao observar a semântica associada às palavras com maior probabilidade dentro do mesmo;
- T_2 : 7,47% **try** + 4,64% **suicide** + 4,24% **go** + 4,02% **keep** + 3,11% **stop** ... – este tópico é ilustrado na Figura 22. As palavras aqui associadas trazem uma ideia de inquietude com o lugar ou estado atual, como observado pela presença de termos

associados ao conceito de movimentação: “**go**”, “**stop**” ou “**fight**” são bons exemplos. Contribui para isto a existência de palavras como “**try**”, “**keep**”, “**attempt**” ou “**think**”, que evocam a ideia de melhorar a própria situação. Sentimentos recorrentes associados a falta de utilidade ou estagnação podem ser indicativos do TDM (ASSOCIATION et al., 2013). Ainda, pode-se perceber termos associados a temática de angústia emocional dentre os listados: “**suicide**”, “**fight**” e “**kill**” são algumas das que vêm à mente, todas essas possuindo importância elevada dentro do tópico;

- T_3 : 4,09% **self** + 3,86% **pain** + 1,8% **fear** + 1,74% **life** + 1,38% **weight** ... – o tópico exibido na Figura 23 situa-se aproximadamente na metade da tabela de tópicos ordenados por coerência. Entretanto, a ideia passada pelo mesmo possui consistência. Isso é notado pela presença de palavras que sugerem o tema de imagem corporal, como “**body**”, “**weight**” e “**self**”, entre outras. Notavelmente, (FLORES-CORNEJO et al., 2017) mostraram que adolescentes insatisfeitos com sua imagem corporal eram mais propensos a demonstrarem sintomas de depressão. Enquanto isso, em (NOLES; CASH; WINSTEAD, 1985) foi constatado que indivíduos com depressão eram mais insatisfeitos com sua imagem corporal e achavam-se menos atraentes que outras pessoas. Ainda, sabe-se que a alteração significativa de peso sem relação com a adoção de dietas é um dos nove sintomas característicos do TDM, segundo o DSM-5 (ASSOCIATION et al., 2013);
- T_4 : 12,27% **love** + 6,04% **girl** + 5,05% **relationship** + 2,93% **guy** + 2,43% **want** ... – este tópico situa-se na média da listagem de tópicos do modelo, e é descrito pela Figura 24. A temática de relacionamentos amorosos pode ser inferida a partir dos termos agregados pelo mesmo. Nota-se a existência de palavras como “**relationship**”, “**girl**”, “**guy**”, “**love**”, esta última sendo a mais importante para o tópico. Relacionamentos interpessoais são de forma geral um assunto delicado para pessoas acometidas com depressão, algo que pode ser mais acentuado caso os relacionamentos em questão tenham um subtexto amoroso. Como exemplo da relevância deste tema para o assunto da depressão, (SHARABI; DELANEY; KNOBLOCH, 2016) estudaram os efeitos da depressão em relacionamentos românticos. Os autores levantaram diferentes categorias de impactos negativos da depressão em relacionamentos, como falta de compreensão entre parceiros, isolamento, falta de energia, dentre outros;
- T_5 : 11,85% **school** + 5,55% **college** + 5,46% **year** + 4,12% **go** + 4,05% **class** ... – a Figura 25 exhibe as palavras associadas a este tópico. Esse é o tópico com maior valor de coerência dentro todos aqueles extraídos pelo modelo explorado neste capítulo. Novamente, o valor da métrica tem amparo nas observações qualitativas a serem feitas sobre o tópico: a ideia de vida estudantil é o cerne do tópico em questão. Enquanto no modelo LDA treinado com o *corpus* em português o tópico

representativo desta temática havia agregado também termos associados ao tema de preocupações financeiras, aqui apenas a rotina de estudos é constituinte do tópico, ampliando a qualidade do mesmo. Como discutido anteriormente, a prevalência de depressão entre estudantes do ensino superior é alta e diversos fatores – como preocupação com o sucesso acadêmico – impactam a saúde mental dos discentes;

- T_6 : 7,84% **make** + 6,96% **know** + 5,29% **feel** + 4,29% **way** + 3,73% **try** ... – o tópico descrito pela Figura 26 encontra-se na parte de baixo da tabela de coerência. De fato, a interpretabilidade do mesmo é difícil, mas algumas de suas palavras trazem uma semântica de introspecção. Termos como “**feel**”, “**think**” e “**know**” têm importância no tópico;
- T_7 : 5,43% **time** + 4,51% **use** + 4,2% **thing** + 3,58% **play** + 3,39% **make** ... – a Figura 27 ilustra este tópico. Apesar da coerência baixa, o tópico inclui palavras associadas ao lazer, *hobbies* e entretenimento. “**Play**”, “**game**”, “**enjoy**”, “**music**”, “**video**” e “**spend**” são alguns dos termos que amparam tal hipótese. A rotina de lazer e *hobbies* é um dos diversos fatores que sofrem o impacto da depressão na vida de um indivíduo, já que pessoas depressivas comumente relatam perda de interesse ou motivação em atividades que antes os agradavam (ASSOCIATION et al., 2013). Deve-se citar, ainda, que manter uma rotina saudável de entretenimento pode auxiliar indivíduos que encontram-se em um estado de humor majoritariamente baixo a recuperarem sua qualidade de vida.

4.2.2.2 Análise qualitativa dos tópicos 8 a 15

Abaixo, os tópicos T_8 a T_{15} são analisados em maiores detalhes.

- T_8 : 8,28% **go** + 6,43% **leave** + 5,62% **move** + 3,79% **live** + 2,57% **place** ... – este tópico é representado pela Figura 28. Palavras como “**leave**”, “**move**”, “**live**” e “**home**” podem ser associadas a uma temática de moradia ou residência, e podem indicar uma discussão sobre independência dos familiares. A saída de casa para ingressar em uma universidade é apontada como um dos estressores da depressão para jovens estudantes, já que frequentemente esta é a primeira vez que os jovens terão de cuidar de suas finanças, comida e residência sem o auxílio dos pais (BEITER et al., 2015);
- T_9 : 21,7% **help** + 15,87% **depression** + 6,86% **need** + 3,82% **problem** + 3,27% **depressed** ... – a Figura 29 ilustra este tópico. Assim como no caso do tópico T_0 , o tópico atual traz uma discussão intimamente associada ao contexto da depressão. Termos como “**help**”, “**depression**”, “**problem**”, “**need**” ou “**struggle**” caracterizam o tópico como referente ao tema das dificuldades atreladas ao transtorno

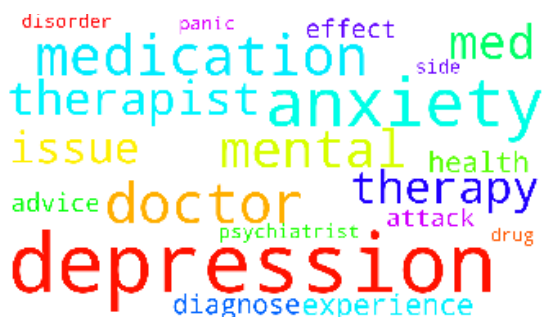


Figura 20 – T_0 do modelo ETM



Figura 21 – T_1 do modelo ETM

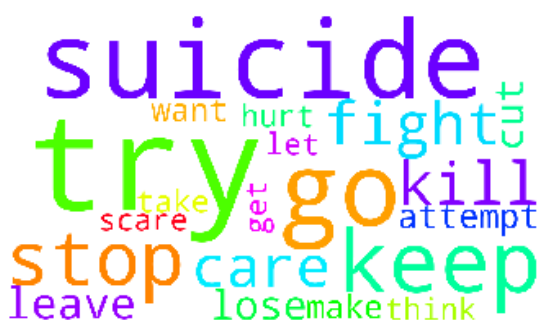


Figura 22 – T_2 do modelo ETM

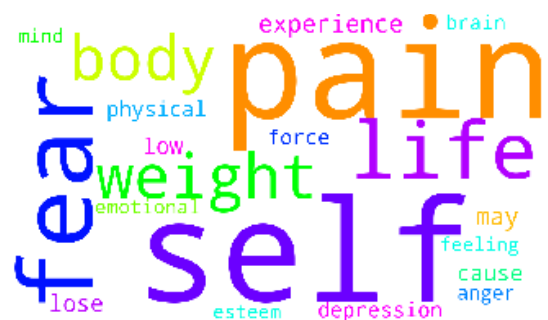


Figura 23 – T_3 do modelo ETM

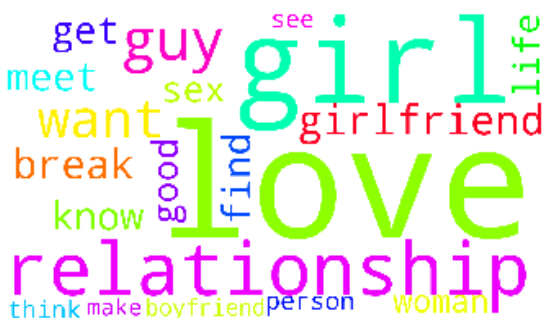


Figura 24 – T_4 do modelo ETM

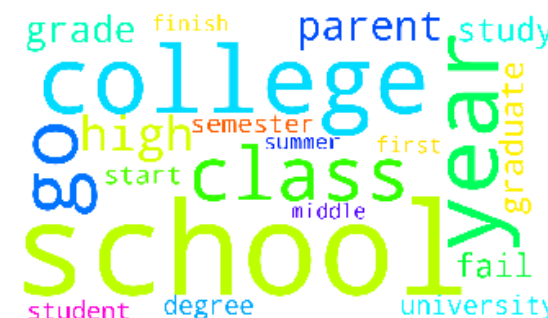


Figura 25 – T_5 do modelo ETM

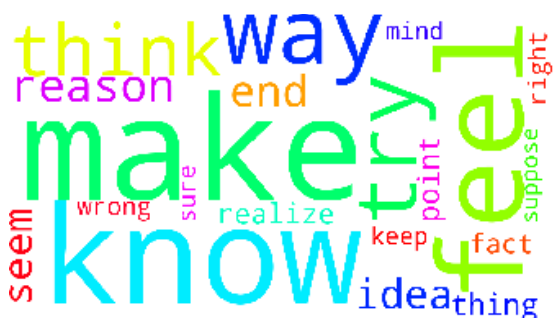


Figura 26 – T_6 do modelo ETM



Figura 27 – T_7 do modelo ETM

depressivo. Nota-se também a presença de palavras associadas à busca por apoio, como “help”, “find”, “support” e “seek”. Ainda, percebe-se que as palavras “help” e “depression” são os termos mais relevantes para este tópico. O primeiro passo

na busca por ajuda costuma ser muito difícil para indivíduos depressivos, muito em razão da auto-estigmatização dessas pessoas e do estigma de pedir ajuda em um problema desse tipo (BARNEY et al., 2006);

- T_{10} : 12,31% **time** + 7,7% **take** + 5,76% **go** + 4,24% **day** + 3,37% **could** ... – este tópico é ilustrado na Figura 30. O tópico associa termos que lembram conceitos temporais, mas é de difícil interpretabilidade, o que tem suporte em seu valor baixo de coerência dentre os tópicos desse modelo. Este seria mais um *junk topic*;
- T_{11} : 25,96% **people** + 4,04% **care** + 3,73% **person** + 3,68% **social** + 2,99% **know** ... – este é um dos tópicos com valor de coerência mais baixo, e é descrito pela Figura 31. Percebe-se a agregação de palavras associadas a sociabilidade no tópico, como “**social**”, “**person**” ou “**people**” – sendo este o termo mais relevante dentro do tópico. O surgimento deste tópico pode estar relacionado às preocupações sociais tipicamente enfrentadas por pessoas depressivas;
- T_{12} : 14,62% **year** + 7,28% **month** + 6,65% **last** + 6,56% **time** + 4,65% **past** ... – assim como o tópico T_{10} , este tópico agrega palavras associadas a conceitos temporais. Pode-se notar isso ao observar a Figura 32. Apesar do tópico ter coerência baixa dentre os avaliados, percebe-se o tema descrito pelo mesmo. A aparição de um tópico como este pode indicar o tema de rotina ou continuidade ao longo do tempo – que é uma das características da depressão;
- T_{13} : 34,25% **want** + 13,33% **know** + 10,3% **go** + 4,47% **feel** + 3,33% **see** ... – o tópico da Figura 33 é caracterizado por sua difícil interpretabilidade. Contudo, palavras indicativas de tensões emocionais podem ser percebidas, como “**cry**”, “**worry**” ou “**feel**”;
- T_{14} : 12,02% **work** + 10,04% **job** + 3,54% **money** + 2,84% **pay** + 2,76% **get** ... – a Figura 34 indica que o tema predominante no tópico é o de vida profissional: palavras como “**work**”, “**job**” e “**money**” são as mais predominantes no tema. Como descrito anteriormente, dificuldades financeiras e expectativas relacionadas ao mercado de trabalho podem ser estressores associados à depressão, sobretudo no caso de estudantes universitários (ANDREWS; WILDING, 2004; BEITER et al., 2015);
- T_{15} : 18,84% **life** + 7,13% **live** + 4,07% **die** + 2,65% **dream** + 2,37% **want** ... – este tópico situa-se na metade de cima da tabela de tópicos e pode ser visto na Figura 35. Aqui encontramos palavras como “**future**”, “**dream**”, “**wish**”, que parecem indicar um anseio por mudanças em algum aspecto da vida por parte do interlocutor. Nota-se uma conotação negativa em diversas palavras, suscitando o tema de emoções negativas, como “**death**”, “**die**” ou “**kill**”. O tópico possui alta

interpretabilidade e caracteriza um humor majoritariamente baixo, característico da depressão.

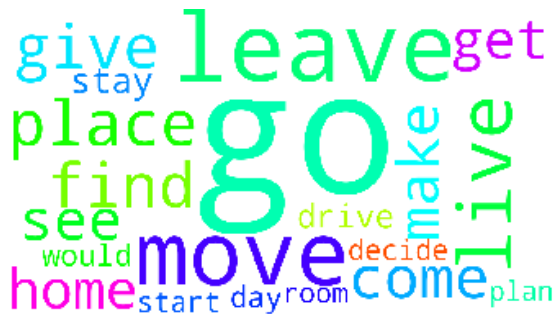


Figura 28 – T_8 do modelo ETM

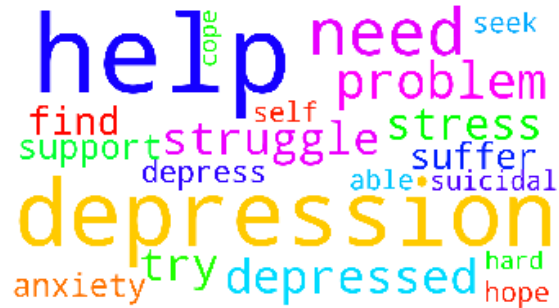


Figura 29 – T_9 do modelo ETM

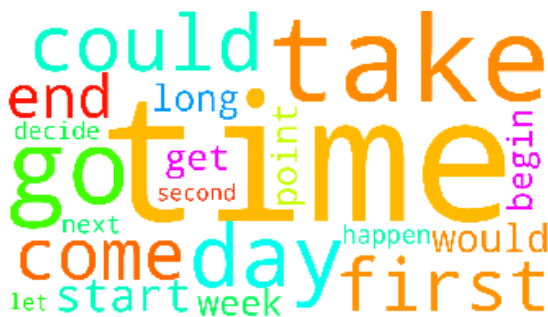


Figura 30 – T_{10} do modelo ETM



Figura 31 – T_{11} do modelo ETM



Figura 32 – T_{12} do modelo ETM

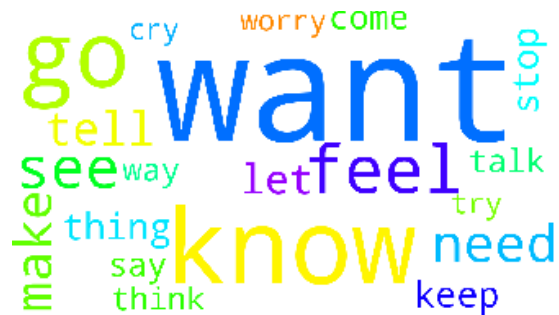


Figura 33 – T_{13} do modelo ETM

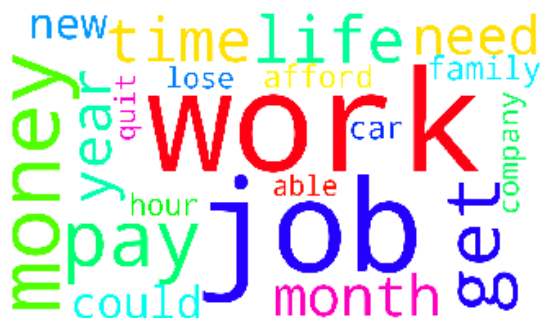


Figura 34 – T_{14} do modelo ETM

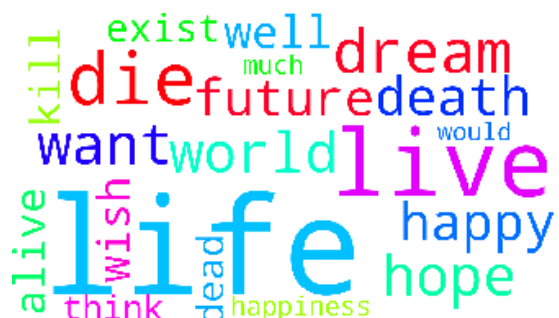


Figura 35 – T_{15} do modelo ETM

4.2.2.3 Análise qualitativa dos tópicos 16 a 23

Prosseguindo com a análise, discute-se abaixo os tópicos de T_{16} a T_{23} .

- T_{16} : 19,42% **would** + 11,18% **think** + 9,07% **could** + 6,45% **know** + 3,31% **wish** ... – o tópico simbolizado pela Figura 36 carrega características de ruminação. Isto se deve ao fato do mesmo incluir palavras que evocam a ideia de reflexão continuada sobre certas situações, como os termos “**could**”, “**would**” e “**think**”, que aparecem com grande importância dentro do mesmo. Deve-se notar, sobretudo, a contribuição enorme da palavra “**would**” para o tópico, chegando a quase 20%, o que caracteriza o raciocínio ancorado em suposições típico da ruminação. Por outro lado, esses são em geral verbos auxiliares na língua inglesa, o que pode ser um indicativo de um tópico com verbos gerais, então não se descarta a possibilidade deste tópico ser um *junk topic*;
- T_{17} : 6,3% **life** + 5,06% **one** + 4,24% **world** + 3,14% **way** + 3,07% **people** ... – a Figura 37 ilustra tal tópico. A agregação de palavras contrasta termos com semântica distinta entre si, como as palavras “**life**” e “**one**” – que são as mais relevantes do tema. Dada a coerência baixa do tópico e sua difícil interpretação, pode-se dizer que o mesmo enquadra-se na definição de *junk topic*;
- T_{18} : 8,07% **talk** + 7,49% **feel** + 5,24% **think** + 4,73% **know** + 2,94% **good** ... – o tópico descrito pela Figura 38 agrega termos associados à interação humana, como “**talk**”, “**thank**” ou “**say**”, mas de forma geral associa palavras que representam ações diversas;
- T_{19} : 46,32% **feel** + 5,86% **happy** + 4,99% **sad** + 4,86% **feeling** + 3,17% **cry** ... – a Figura 39 ilustra este tópico, que tem como termo predominante a palavra “**feel**”. A probabilidade desta palavra no tópico chega a quase 50%, caracterizando a temática de sentimentos. Outros termos com importância relevante são “**happy**”, “**sad**”, “**feeling**”, “**cry**”, “**lonely**”, entre outros. Deve-se notar que a maioria dos termos associados tem valoração negativa, o que confere um humor negativo ao tópico. Ainda, percebe-se que o tópico está situado na parte superior da tabela de coerência;
- T_{20} : 14,71% **get** + 10,38% **go** + 7,93% **bad** + 7,29% **start** + 7,07% **thing** ... – este tópico é representado na Figura 40. O conjunto de termos descrito por este tópico é bastante difícil de interpretar, o que constata o fato do mesmo ter o menor valor de coerência. Este é mais um exemplo de *junk topic*;
- T_{21} : 3,03% **look** + 2,68% **see** + 2,2% **head** + 2,0% **cry** + 1,85% **walk** ... – o tópico ilustrado na Figura 41 abarca palavras contrastantes. Termos como “**look**” e “**remember**” ou “**head**” e “**room**”

- T_{22} : 4,54% **parent** + 4,37% **family** + 4,07% **year** + 3,98% **mom** + 3,12% **mother** ... – percebe-se no tópico exibido na Figura 42 a predominância da temática de família. O tópico, que possui o terceiro maior valor de coerência, associa termos como “**parent**”, “**family**” e “**mom**”, conferindo grande importância aos mesmos dentro do tópico. Percebe-se novamente a importância do assunto vida familiar para o tema depressão, já que este tópico apareceu também na observação dos resultados oriundos do *corpus* em língua portuguesa. Deve-se reforçar que ambientes familiares instáveis podem ser um contribuidor para o desenvolvimento de depressão em jovens (WANG et al., 2020);
- T_{23} : 11,22% **day** + 6,88% **sleep** + 5,68% **go** + 4,22% **night** + 4,21% **bed** ... – a Figura 43 lista os termos mais relevantes para o tópico, que é um daqueles com maior valor para coerência. Nota-se uma predominância de palavras que evocam rotina diária de afazeres neste tópico, como por exemplo “**day**”, “**sleep**”, “**go**”, “**bed**” e “**wake**”. Sabe-se que a depressão pode afetar extensivamente a rotina diária de um paciente, apresentando impactos nos mais diversos âmbitos de sua vida.

4.2.2.4 Análise qualitativa dos tópicos 24 a 17

Por fim, os tópicos restantes são explorados a seguir.

- T_{24} : 13,39% **say** + 11,63% **tell** + 6,77% **ask** + 4,73% **talk** + 4,31% **call** ... – o tópico representado na Figura 44 inclui múltiplas palavras que evocam o tema de conversações humanas. O tema é percebido pela presença de termos como “**say**”, “**tell**” e “**call**”, entre outros. A co-ocorrência dessas palavras justifica o valor de coerência do tópico, que situa o mesmo na metade de cima da tabela de tópicos do modelo. Como dito anteriormente, dificuldades em relações interpessoais costumam ser um problema citado por pessoas depressivas, e o presente tópico ilustra isso;
- T_{25} : 26,7% **friend** + 5,93% **good** + 4,97% **talk** + 4,62% **go** + 3,82% **time** ... – a Figura 45 ilustra este tópico. A palavra com maior importância no tópico é “**friend**”, com mais de 26% de probabilidade. De fato, o tema amizade é o predominante neste tópico, já que o mesmo inclui palavras associadas à interação com amigos, como “**talk**”, “**hang**”, “**party**”, “**group**”, entre outras. Contudo, o tópico também inclui as palavras “**girlfriend**” e “**family**”, indicando que o mesmo abrange outros âmbitos de relacionamento;
- T_{26} : 5,3% **work** + 3,87% **thing** + 3,83% **time** + 3,37% **try** + 3,03% **find** ... – este tópico é descrito pela Figura 46 e o mesmo abarca palavras com amplo significado. Algumas palavras evocam uma temática de reflexão e introspecção, semelhante à vista no tópico T_{16} . Porém, o tópico tem baixa interpretabilidade, o que é atestado

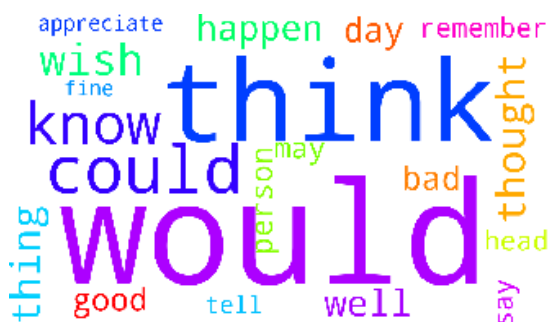


Figura 36 – T_{16} do modelo ETM

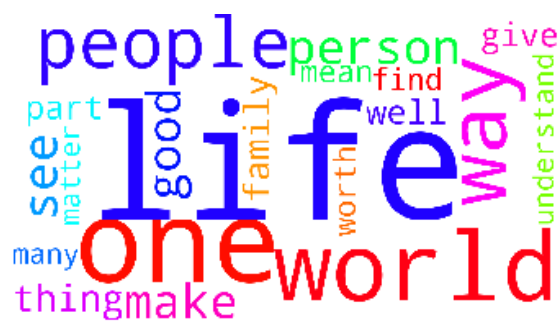


Figura 37 – T_{17} do modelo ETM



Figura 38 – T_{18} do modelo ETM

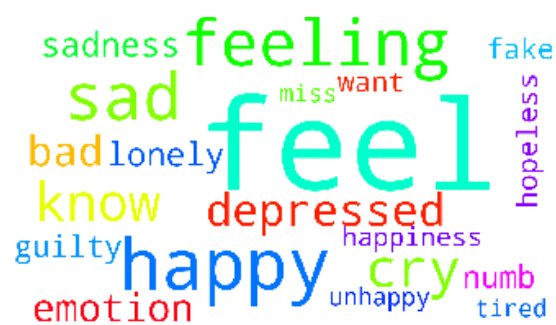


Figura 39 – T_{19} do modelo ETM



Figura 40 – T_{20} do modelo ETM



Figura 41 – T_{21} do modelo ETM



Figura 42 – T_{22} do modelo ETM

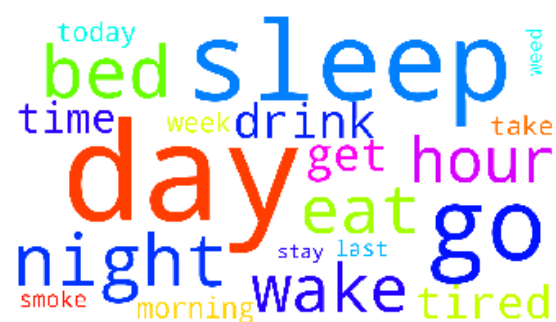
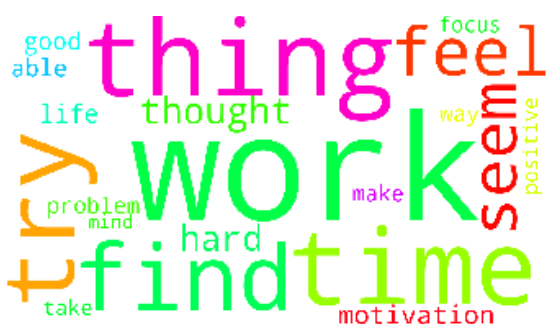
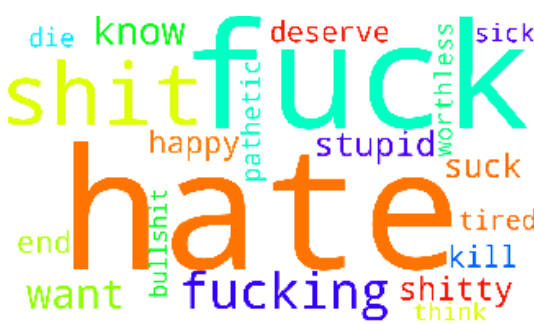


Figura 43 – T_{23} do modelo ETM

pelo fato do mesmo estar situado na metade inferior da tabela de coerência. Dessa forma, é difícil identificar seu tema predominante;

- T_{27} : 10,86% **hate** + 10,59% **fuck** + 8,34% **shit** + 5,22% **fucking** + 3,38% **want** ... – o tópico ilustrado pela Figura 47 encontra-se na metade superior da tabela de coerência. Pode-se perceber que, dentre as palavras com maior probabilidade dentro do tópico, a maioria delas é caracterizada por termos com semântica negativa associada. Palavras como “**hate**”, “**fuck**”, “**shit**”, entre outras dominam o tópico. Sendo assim, o tópico pode ser caracterizado como representativo de negatividade.

Figura 44 – T_{24} do modelo ETMFigura 45 – T_{25} do modelo ETMFigura 46 – T_{26} do modelo ETMFigura 47 – T_{27} do modelo ETM

4.2.3 Análise léxica

A Figura 48 exibe as cinco categorias lexicais mais predominantes entre os tópicos extraídos com o modelo ETM com $K = 28$ tópicos. As categorias foram obtidas a partir da biblioteca *empath-client*. Apenas as categorias-padrão do *Empath* foram consideradas, sem o emprego da criação dinâmica de novas categorias possibilitada pela ferramenta. Novamente, os vinte termos mais relevantes dentro de cada tópico foram considerados para a obtenção das categorias, e assim como o LIWC, o *Empath* também pode categorizar uma mesma palavra em diferentes categorias lexicais. Nota-se sobretudo na Figura a predominância de categorias com valoração semântica negativa, como *negative_emotion*, *pain*, *sadness* e *suffering*.

Segundo o *Empath*, 143 categorias distintas podem ser observadas no conjunto de tópicos avaliado. Associada à mais de 8% das palavras analisadas, a categoria *ne-*

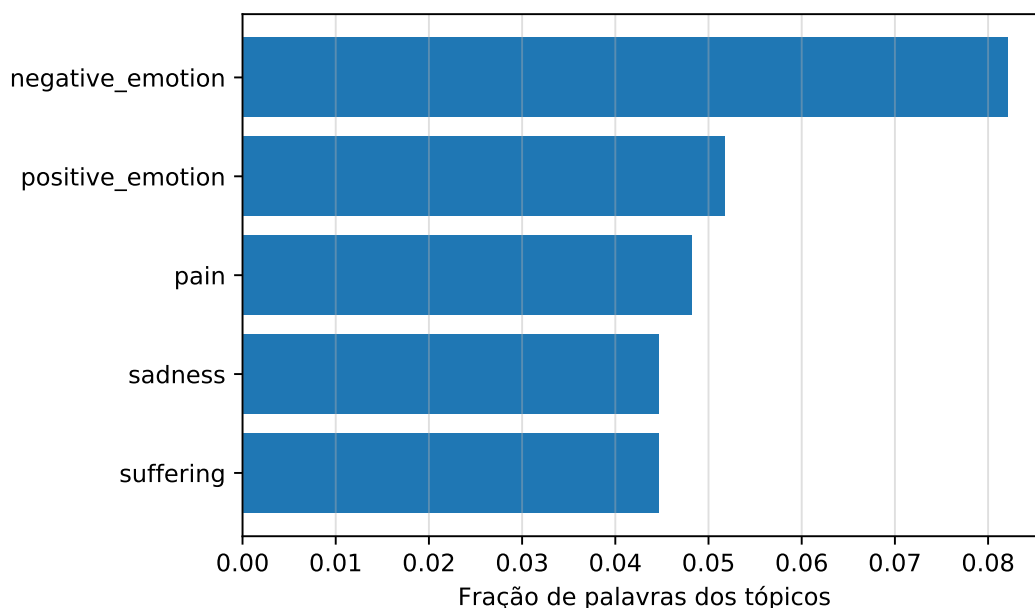


Figura 48 – Cinco categorias lexicais predominantes nos tópicos do modelo ETM com $K = 28$ treinado em inglês, segundo o *Empath*.

negative_emotion predomina de forma isolada em relação às demais. Como exemplos de palavras associadas à mesma, pode-se citar “fight”, “hurt”, “kill”, “pain”, “depressed”, “cry” e “wrong”, entre outras. A categorização tem clara significância, já que essas palavras aparecem em diversos dos tópicos analisados. Os tópicos T_{15} , T_{19} e T_{27} são alguns dos quais mais se relacionam com esta categoria, agregando de forma predominante palavras que evocam emoções negativas e negatividade em geral. Em seguida, a segunda categoria mais predominante é a de *positive_emotion*. Palavras como “care”, “love”, “hope” e “happy” foram algumas daquelas associadas à categoria. Vários tópicos agregam palavras desse tipo, em número minoritário, como os tópicos T_4 – associado a relacionamentos amorosos – e T_{15} – associado a anseios por mudança pessoal. Em seguida, são listadas mais três categorias intimamente associadas à semântica negativa: *pain*, *sadness* e *suffering*. As palavras “kill” e “cry” é exemplo das categorias *pain* e *suffering*. “Cry” também está associado à categoria *sadness*, assim como a palavra de mesmo nome. O tópico T_{19} é um dos mais representativos dessas categorias. Portanto, pode-se perceber que as três categorias estão bastante entrelaçadas entre si. A alta predominância da temática indica a importância deste tema dentro do *corpus* explorado, fortalecendo as ligações entre os tópicos latentes e o tema da depressão.

4.3 Discussão geral

Diante do discutido acima, constata-se que tópicos latentes relevantes para a discussão da depressão foram encontrados em ambos os casos. Notavelmente, tópicos associados

a temas como família, relacionamentos interpessoais e vida estudantil foram encontrados nos dois modelos analisados, tanto em português quanto em inglês. A importância desse resultado reside no fato de que, apesar das amplas diferenças na constituição de ambos os *corpus*, temáticas semelhantes puderam ser observadas. Entre as diferenças mais significativas entre os conjuntos textuais, pode-se citar que o *corpus* de língua portuguesa foi construído a partir de postagens coletadas de *subreddits* não específicos ao tema depressão, enquanto um *subreddit* focado no tema foi usado para a construção do *corpus* em inglês. Além disso, deve-se citar a diferença de proporção entre ambos os *corpus*: o conjunto textual trabalhado em inglês possui quase dez vezes mais postagens do que sua contraparte em português. Ainda, deve-se citar a óbvia diferença linguística entre ambos os conjuntos, já que as particularidades de cada idioma também afetam os resultados obtidos por meio da modelagem de tópicos aplicada aos mesmos.

Nesse contexto, a diferença linguística fica acentuada em razão da necessidade de uma etapa de filtragem de vocabulário específica para cada idioma analisado. Outro fator a apoiar tal percepção reside nas diferenças entre as categorias de POS utilizadas no treinamento realizado para cada idioma. Enquanto resultados satisfatórios foram obtidos com o *corpus* inglês ao manter-se categorias como verbos, substantivos e adjetivos, o mesmo não foi observado no *corpus* português. Para este, foi preciso expandir a filtragem de categorias, mantendo na versão final analisada apenas as palavras que representavam substantivos. Além disso, houve também diferenças em termos de tópicos extraídos pelos modelos, já que o modelo treinado no *corpus* em inglês conseguiu determinar um número bem mais amplo de temas latentes que sua contraparte LDA treinada em português. Em inglês, tópicos específicos para temas como relacionamentos amorosos, vida profissional e negatividade são notados, algo que não foi percebido para o *corpus* em português.

Considerando os tópicos extraídos, percebe-se que aqueles observados no modelo treinado em inglês tiveram maiores níveis de coerência em relação aos obtidos pelo modelo treinado em português. A interpretabilidade dos tópicos em inglês foi mais facilitada em razão disso, o que reforça a correlação empreendida entre a métrica de coerência NPMI e o julgamento humano. Enquanto tópicos do modelo em português agregam com maior frequência palavras com semântica difusa entre si, isso acontece de forma bastante inferior quando analisa-se os tópicos em inglês. Outra característica percebida na análise qualitativa dos tópicos em ambos os idiomas foi a predominância ou não dos chamados *junk topics*, que são tópicos com pouca semântica interna associada. No modelo LDA com $K = 5$ percebe-se a presença de um *junk topic*, enquanto os demais tópicos exprimem uma ideia predominante, ainda que não tão assertiva se comparados ao modelo treinado em inglês. Enquanto isso, o modelo ETM com $K = 28$ apresentou mais de vinte tópicos interpretáveis, muitos dos quais associados à discussão da depressão. Dessa forma, pode-se citar uma menor predominância de tópicos de baixa interpretabilidade no modelo avaliado em inglês, guardadas as devidas proporções entre os modelos relativas ao número de tópicos

e tamanho do *corpus* explorado em cada cenário. Por fim, a análise léxica de ambos os modelos ilustrou a melhor capacidade de apreensão de temáticas associadas à depressão demonstrada pelo modelo ETM, constatando a relação entre os termos agregados nos tópicos com categorias lexicais de valoração semântica predominantemente negativa. Tal associação não foi percebida de forma clara nos tópicos extraídos em português pelo modelo LDA. No caso do modelo em português, os resultados em termos de coerência e interpretabilidade podem ser atribuídos tanto à menor capacidade da arquitetura LDA em extrair relações de contexto semântico dos conjuntos textuais, como pela não-especificidade do *corpus* usado, quanto pela proporção reduzida do mesmo. Algoritmos de processamento textual baseados em modelagens estatísticas são beneficiados pelo uso de grandes conjuntos de dados, o que não é diferente para a modelagem de tópicos (ALLISON; GUTHRIE; GUTHRIE, 2006).

Discutindo os resultados por tipo de arquitetura para modelagem de tópicos, a característica principal é o melhor desempenho demonstrado pelos modelos ETM na maioria dos casos. Apesar do ETM não ter sido o modelo com melhor coerência em português, ele foi o melhor modelo na maioria dos cenários explorados para o idioma em termos de número de tópicos K . Desempenho semelhante do modelo foi observado nos resultados de treinamento baseados em inglês, onde o ETM predominou em todos os cenários, exceto um. Por outro lado, os resultados demonstrados pelos modelos CTM não foram congruentes com sua capacidade de identificação de relações contextuais entre palavras, baseada no SBERT. Em razão da necessidade de realização de etapas de pré-processamento semelhantes para os três tipos de modelos de forma a unificar seu treinamento, é possível que características do *corpus* que eram relevantes para o CTM tenham sido removidas, prejudicando seu treinamento. Uma análise com maior foco nesta arquitetura para modelagem de tópicos seria capaz de elucidar a validade desta percepção.

5 Conclusão

A presente monografia buscou explorar métodos de modelagem de tópicos para identificação de temas latentes relacionados à depressão no âmbito da rede social *Reddit*. Diferentes técnicas para modelagem de tópicos existem na literatura, e três foram escolhidas para maior aprofundamento neste trabalho: LDA, CTM e ETM. Enquanto a primeira tem uma arquitetura baseada exclusivamente em métodos probabilísticos para extração de tópicos, as demais utilizam o conceito de *embeddings* de palavras para depreender maior semântica dos tópicos observados em um *corpus*. Adicionalmente, buscou-se validar as observações oriundas da modelagem de tópicos por meio da análise léxica de categorias de palavras presentes nos tópicos latentes. Para tal, as ferramentas LIWC e *Empath* foram utilizadas, ampliando o entendimento dos resultados obtidos por meio da modelagem de tópicos.

A depressão é um problema humano, não restrito às barreiras impostas por países, continentes ou idiomas. Consequentemente, o presente estudo buscou como principal objetivo identificar as possíveis similaridades e distinções entre as discussões associadas à depressão em diferentes linguagens. Para o estudo, um *corpus* de postagens do *Reddit* relativo ao tema foi coletado nas línguas portuguesa e inglesa. As diferentes características dos idiomas impactaram a metodologia proposta no trabalho, já que algumas das etapas tiveram de ser adaptadas às especificidades de cada linguagem. Ainda, deve-se notar que enquanto um *subreddit* intitulado *depression* foi usado para construção do *corpus* em inglês, não foi possível realizar o análogo para o *corpus* em português. Diante disso, *subreddits* genéricos em português foram usados para construção do *corpus* para o idioma, mediante a busca por palavras-chave associadas à depressão. Ainda, a variedade de postagens associadas ao tema encontradas em inglês e português teve enorme variação, causando uma diferença de dez vezes mais postagens coletadas em inglês.

Entretanto, as diferenças entre as linguagens trabalhadas e na constituição de cada *corpus* não impediram que padrões de similaridade fossem percebidos. Tópicos semelhantes foram percebidos nos modelos treinados em ambos os idiomas, com variações nos níveis de especificidade e qualidade das agregações. Sendo assim, foi possível identificar algumas temáticas que surgem em discussões sobre depressão em ambos os idiomas. Deve-se notar ainda que, em português, temas relacionados à depressão foram percebidos mesmo em um *corpus* construído por postagens de comunidades generalistas, não específicas ao tema. Além disso, foi possível perceber também o impacto da especificidade e proporção de cada *corpus* nos resultados. Os tópicos encontrados em inglês foram mais assertivos, coerentes e mais correlacionados à depressão segundo a análise de categorias lexicais. Os tópicos em português foram mais prejudicados, principalmente em razão de suas características de

constituição discutidas anteriormente. Ainda, os tópicos em inglês tiveram maior variedade temática, agregando temas que não apareceram nos tópicos extraídos pelos modelos treinados no reduzido *corpus* de língua portuguesa.

Pode-se dizer, ainda, que os resultados obtidos serviram para reforçar a aplicabilidade da modelagem de tópicos para identificação de temáticas em redes sociais, já que tópicos relevantes para o tema discutido neste estudo foram observados em diferentes tipos de modelos. Novamente, é preciso ressaltar que, mesmo com uma grande disparidade em proporção e constituição dos *corpora*, foi possível observar temas relevantes para discussões sobre depressão, em ambos os idiomas explorados. Nesse contexto, especial atenção é devida aos modelos ETM, que tiveram predominância entre os melhores modelos observados tanto em português, quanto em inglês. Os resultados ampliam as possibilidades de uso das técnicas de NLP como auxílio a estratégias que possam vir a estudar, identificar e intervir em potenciais casos de depressão percebidos nas redes sociais.

Contudo, também podem ser enumeradas ameaças à validade dos resultados apresentados. A primeira delas diz respeito à análise de categorias lexicais realizada. Devido à incompatibilidade entre a ferramenta usada para extração de categorias LIWC e o dicionário em português mais recente no *software*, foi preciso utilizar uma versão antiga do dicionário para o presente trabalho. Versões mais recentes do dicionário poderiam proporcionar uma análise léxica mais aprofundada. Além disso, devido a limitações impostas pelo próprio LIWC, não foi possível utilizar seus dicionários para análise dos tópicos em ambos os idiomas aqui explorados. Apesar do *Empath* possuir alta correlação em termos de categorias com o próprio LIWC, uma comparação entre os achados obtidos por ambas as ferramentas fica prejudicada já que nem todas as categorias existentes em uma ferramenta possuem análogas na outra.

Em relação às arquiteturas de modelagem de tópicos exploradas nesta monografia, também deve-se citar o baixo desempenho percebido nos resultados de treinamento para os modelos CTM. Os resultados observados contradizem a percepção de que o modelo teria melhor capacidade de percepção das relações contextuais de palavras em cada *corpus*. Contudo, isso pode representar uma falha na provisão de recursos para o treinamento desse tipo de modelo. A necessidade de utilizar entradas semelhantes para cada tipo de modelo de tópicos pode ter resultado na perda de características caras ao CTM em seu processo de modelagem. Apenas uma análise focada na arquitetura CTM poderá identificar precisamente as razões para os resultados observados aqui.

O presente trabalho também possibilita caminhos futuros para a pesquisa. Em relação à análise léxica realizada, no futuro o uso de uma ferramenta única contribuiria para um aprimoramento da referida avaliação. Em relação aos treinamentos de modelos de tópicos, os modelos baseados em *embeddings* de palavras poderiam beneficiar-se do emprego de *embeddings* pré-treinados em *corpora* de depressão. O presente trabalho indica que a

construção de *embeddings* desse tipo pode ser realizada por meio da coleta de postagens associadas a discussões sobre depressão no *Reddit*. Dessa forma, tópicos intrinsecamente associados à temática poderiam ser extraídos com melhor qualidade. Além disso, na falta de *subreddits* específicos ao tema, o uso de tal conjunto de dados possibilitaria a filtragem de temas associados à depressão em português na rede social. Outra abordagem possível seria explorar as possibilidades de treinamento multilíngue proporcionadas pelo CTM em sua completude: resultados relevantes poderiam ser obtidos treinando-se o modelo CTM com um *embedding* de determinada linguagem e inferindo os tópicos escritos em outra. Tal abordagem alinha-se com os objetivos deste trabalho, já que possibilita de forma direta a aplicação de tópicos aprendidos em um idioma para entendimento de um *corpus* escrito em outro.

Uma vez discutidos os pontos acima, é preciso citar a contribuição do estudo para o melhor entendimento do problema da depressão e para a melhor compreensão e aprendizado de um subconjunto de técnicas disponibilizadas pela NLP para análise textual. O presente estudo abordou os pré-requisitos para uma modelagem de tópicos de qualidade, mostrou o impacto das diferenças linguísticas em um contexto de modelagem em idiomas distintos e ilustrou a viabilidade das técnicas exploradas dentro do âmbito das redes sociais. O objetivo do trabalho foi cumprido, e os resultados apresentados possibilitam o desenvolvimento de novos estudos, de forma a aprimorar o entendimento sobre os achados aqui descritos.

Referências

- AL-MOSAIWI, M.; JOHNSTONE, T. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, Sage Publications Sage CA: Los Angeles, CA, v. 6, n. 4, p. 529–542, 2018. Citado 3 vezes nas páginas 2, 3 e 30.
- ALAMMAR, J. *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)*. 2018. Acesso em: 19 dez. 2020. Disponível em: <<http://jalamar.github.io/illustrated-bert/>>. Citado na página 17.
- ALGHAMDI, R.; ALFALQI, K. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, Citeseer, v. 6, n. 1, 2015. Citado na página 18.
- ALLISON, B.; GUTHRIE, D.; GUTHRIE, L. Another look at the data sparsity problem. In: SPRINGER. *International Conference on Text, Speech and Dialogue*. [S.l.], 2006. p. 327–334. Citado 2 vezes nas páginas 9 e 68.
- ALSUMAIT, L. et al. Topic significance ranking of lda generative models. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2009. p. 67–82. Citado na página 49.
- ANDREWS, B.; WILDING, J. M. The relation of depression and anxiety to life-stress and achievement in students. *British journal of psychology*, Wiley Online Library, v. 95, n. 4, p. 509–521, 2004. Citado na página 60.
- ASSOCIATION, A. P. et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. [S.l.]: American Psychiatric Pub, 2013. Citado 3 vezes nas páginas 30, 57 e 58.
- ASSUNÇÃO, A. L. de. *Variação lingüística, uma realidade de nossa língua*. 2010. Acesso em: 19 dez. 2020. Disponível em: <<https://monografias.brasilecola.uol.com.br/educacao/variacao-linguistica-uma-realidade-nossa-lingua.htm>>. Citado na página 7.
- ASUNCION, A. et al. On smoothing and inference for topic models. *arXiv preprint arXiv:1205.2662*, 2012. Citado na página 22.
- BALDOMINOS, A.; SAEZ, Y.; ISASI, P. Evolutionary convolutional neural networks: An application to handwriting recognition. *Neurocomputing*, Elsevier, v. 283, p. 38–52, 2018. Citado na página 11.
- BARNEY, L. J. et al. Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, Sage Publications Sage UK: London, England, v. 40, n. 1, p. 51–54, 2006. Citado na página 60.
- BARRY, C. T. et al. Adolescent social media use and mental health from adolescent and parent perspectives. *Journal of adolescence*, Elsevier, v. 61, p. 1–11, 2017. Citado na página 1.

- BARUA, J. *Word Embeddings Versus Bag-of-Words: The Curious Case of Recommender Systems*. 2020. Acesso em: 09 jan. 2021. Disponível em: <<https://medium.com/swlh/word-embeddings-versus-bag-of-words-the-curious-case-of-recommender-systems-6ac1604d4424>>. Citado 2 vezes nas páginas 8 e 10.
- BAUMGARTNER, J. et al. The pushshift reddit dataset. In: *Proceedings of the International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2020. v. 14, p. 830–839. Citado na página 34.
- BEITER, R. et al. The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of affective disorders*, Elsevier, v. 173, p. 90–96, 2015. Citado 3 vezes nas páginas 50, 58 e 60.
- BENGIO, Y. et al. A neural probabilistic language model. *The journal of machine learning research*, JMLR. org, v. 3, p. 1137–1155, 2003. Citado na página 10.
- BIANCHI, F.; TERRAGNI, S.; HOVY, D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*, 2020. Citado 3 vezes nas páginas 19, 23 e 25.
- BIANCHI, F. et al. Cross-lingual contextualized topic models with zero-shot learning. In: *EACL*. [S.l.: s.n.], 2021. Citado 3 vezes nas páginas 5, 19 e 23.
- BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. [S.l.]: "O'Reilly Media, Inc.", 2009. Citado na página 37.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado 4 vezes nas páginas 5, 19, 20 e 22.
- BOYD-GRABER, J. L. et al. *Applications of topic models*. [S.l.]: now Publishers Incorporated, 2017. v. 11. Citado na página 21.
- BROWN, C. *liwc-python*. 2012. Acesso em: 15 abr. 2021. Disponível em: <<https://github.com/chbrown/liwc-python>>. Citado na página 45.
- BROWN, T. B. et al. *Language Models are Few-Shot Learners*. 2020. Citado na página 11.
- BROWNLE, J. *A Gentle Introduction to the Bag-of-Words Model*. 2017. Acesso em: 06 dez. 2020. Disponível em: <<https://machinelearningmastery.com/gentle-introduction-bag-words-model/>>. Citado na página 9.
- BROWNLE, J. *What Are Word Embeddings for Text?* 2017. Acesso em: 06 dez. 2020. Disponível em: <<https://machinelearningmastery.com/what-are-word-embeddings/>>. Citado na página 12.
- BURKOV, A. *Word Embeddings & Self-Supervised Learning, Explained*. 2019. Acesso em: 17 fev. 2021. Disponível em: <<https://www.kdnuggets.com/2019/01/burkov-self-supervised-learning-word-embeddings.html>>. Citado na página 13.
- CAO, L.; FEI-FEI, L. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: IEEE. *2007 IEEE 11th International Conference on Computer Vision*. [S.l.], 2007. p. 1–8. Citado na página 19.

- CAO, X. et al. Understanding the influence of social media in the workplace: An integration of media synchronicity and social capital theories. In: IEEE. *2012 45th Hawaii International Conference on System Sciences*. [S.l.], 2012. p. 3938–3947. Citado na página 1.
- CHOUDHURY, M. D.; DE, S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In: *Eighth international AAAI conference on weblogs and social media*. [S.l.: s.n.], 2014. Citado na página 2.
- CHUANG, J. et al. Large-scale examination of academic publications using statistical models. In: *Proc. AVI Workshop Supporting Asynchronous Collaboration Visual Anal. Syst.* [S.l.: s.n.], 2012. Citado na página 19.
- CLARK, J.; KOPRINSKA, I.; POON, J. A neural network based approach to automated e-mail classification. In: IEEE. *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*. [S.l.], 2003. p. 702–705. Citado na página 11.
- CRUNCHBASE. *Reddit - Crunchbase*. 2021. Acesso em: 01 mar. 2021. Disponível em: <<https://www.crunchbase.com/organization/reddit>>. Citado na página 32.
- CURISKIS, S. A. et al. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, Elsevier, v. 57, n. 2, p. 102034, 2020. Citado na página 18.
- DATA, O. W. I. *Number of people using social media platforms, 2004 to 2019*. 2020. Acesso em: 06 dez. 2020. Disponível em: <<https://ourworldindata.org/grapher/users-by-social-media-platform>>. Citado 3 vezes nas páginas ix, 1 e 2.
- DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Citado 4 vezes nas páginas ix, 12, 15 e 16.
- DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. [S.l.: s.n.], 2019. p. 4171–4186. Citado 3 vezes nas páginas 8, 11 e 14.
- DEVOPEDIA. *BERT (Language Model)*. 2019. Acesso em: 20 fev. 2021. Disponível em: <<https://devopedia.org/bert-language-model>>. Citado na página 18.
- DIENG, A. B.; RUIZ, F. J.; BLEI, D. M. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 8, p. 439–453, 2020. Citado 4 vezes nas páginas 5, 20, 25 e 26.
- EICHSTAEDT, J. C. et al. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 115, n. 44, p. 11203–11208, 2018. Citado 3 vezes nas páginas 3, 28 e 29.
- FALLER, A. *What is Natural Language Processing and Why is it Hard?* 2020. Acesso em: 05 jan. 2021. Disponível em: <<https://www.colorado.edu/earthlab/2020/02/07/what-natural-language-processing-and-why-it-hard>>. Citado na página 7.

- FAST, E. *empath-client*. 2016. Acesso em: 17 abr. 2021. Disponível em: <<https://github.com/Ejhf/empath-client>>. Citado na página 45.
- FAST, E.; CHEN, B.; BERNSTEIN, M. S. Empath: Understanding topic signals in large-scale text. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. [S.l.: s.n.], 2016. p. 4647–4657. Citado na página 28.
- FERNANDES, M. A. et al. Prevalence of anxious and depressive symptoms in college students of a public institution. *Revista brasileira de enfermagem*, SciELO Brasil, v. 71, p. 2169–2175, 2018. Citado na página 50.
- FILHO, P. B.; PARDO, T. A. S.; ALUÍSIO, S. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2013. Citado 2 vezes nas páginas 28 e 51.
- FIRTH, J. R. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, Basil Blackwell, 1957. Citado na página 10.
- FLORES-CORNEJO, F. et al. Association between body image dissatisfaction and depressive symptoms in adolescents. *Brazilian Journal of Psychiatry*, SciELO Brasil, v. 39, n. 4, p. 316–322, 2017. Citado na página 57.
- GAUR, M.; KURSUNCU, U.; ALAMBO, A. "let me tell you about your mental health!" contextualized classification of reddit posts to dsm-5 for web-based intervention. In: *CIKM*. [S.l.: s.n.], 2018. p. 753–762. Citado na página 30.
- GILBERT, E. Widespread underprovision on reddit. In: *Proceedings of the 2013 conference on Computer supported cooperative work*. [S.l.: s.n.], 2013. p. 803–808. Citado na página 32.
- GKOTSIS, G. et al. The language of mental health problems in social media. In: *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. [S.l.: s.n.], 2016. p. 63–73. Citado na página 2.
- GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. *Proceedings of the National academy of Sciences*, National Acad Sciences, v. 101, n. suppl 1, p. 5228–5235, 2004. Citado na página 22.
- HARRIS, Z. S. Distributional structure. *Word*, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954. Citado na página 8.
- HINTON, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation*, MIT Press, v. 14, n. 8, p. 1771–1800, 2002. Citado na página 24.
- HOFFMAN, M.; BACH, F. R.; BLEI, D. M. Online learning for latent dirichlet allocation. In: *advances in neural information processing systems*. [S.l.: s.n.], 2010. p. 856–864. Citado na página 22.
- HONNIBAL, M.; MONTANI, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. 2017. Citado na página 36.

- HOREV, R. *BERT Explained: State of the art language model for NLP*. 2018. Acesso em: 19 dez. 2020. Disponível em: <<https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>>. Citado na página 17.
- HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. Citado na página 15.
- KAFKA, F. *O processo*. [S.l.]: Editora Vozes, 2019. Citado 2 vezes nas páginas ix e 20.
- KANDEL, D. B.; RAVEIS, V. H.; DAVIES, M. Suicidal ideation in adolescence: Depression, substance use, and other risk factors. *Journal of Youth and Adolescence*, Springer, v. 20, n. 2, p. 289–309, 1991. Citado na página 1.
- KARLSSON, V. *SentenceBERT — Semantically meaningful sentence embeddings the right way*. 2020. Acesso em: 20 fev. 2021. Disponível em: <<https://medium.com/dair-ai/tl-dr-sentencebert-8dec326daf4e>>. Citado na página 18.
- KECHAGIA, M. *What is topic modeling?* 2015. Acesso em: 29 dez. 2020. Disponível em: <<https://blog.xrds.acm.org/2015/07/what-is-topic-modeling/>>. Citado na página 18.
- KELLY, Y. et al. Social media use and adolescent mental health: Findings from the uk millennium cohort study. *EClinicalMedicine*, Elsevier, v. 6, p. 59–68, 2018. Citado na página 1.
- LAI, S. et al. Recurrent convolutional neural networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 29, n. 1, Feb. 2015. Disponível em: <<https://ojs.aaai.org/index.php/AAAI/article/view/9513>>. Citado na página 11.
- LAU, J. H.; NEWMAN, D.; BALDWIN, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 2014. p. 530–539. Citado na página 43.
- LEE, R. S. Natural language processing. In: *Artificial Intelligence in Daily Life*. [S.l.]: Springer, 2020. p. 157–192. Citado na página 7.
- LIDDY, E. D. Natural language processing. 2001. Citado na página 7.
- LIU, L. et al. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, Springer, v. 5, n. 1, p. 1608, 2016. Citado na página 19.
- LIU, Y. et al. A crowdsourcing-based topic model for service matchmaking in internet of things. *Future Generation Computer Systems*, Elsevier, v. 87, p. 186–197, 2018. Citado na página 19.
- LIU, Z.; LIN, Y.; SUN, M. Representation learning and nlp. In: _____. *Representation Learning for Natural Language Processing*. Singapore: Springer Singapore, 2020. p. 1–11. ISBN 978-981-15-5573-2. Disponível em: <https://doi.org/10.1007/978-981-15-5573-2_1>. Citado 3 vezes nas páginas 8, 10 e 12.
- LUTINS, E. *Grid Searching in Machine Learning: Quick Explanation and Python Implementation*. 2017. Acesso em: 18 fev. 2021. Disponível em: <<https://elutins.medium.com/grid-searching-in-machine-learning-quick-explanation-and-python-implementation-550552200596>>. Citado na página 23.

- MANNING, C.; SCHUTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT press, 1999. Citado na página 3.
- MARCHAL, F. *File:Nokota Horses.jpg*. 2005. Acesso em: 21 dez. 2020. Disponível em: <https://commons.wikimedia.org/wiki/File:Nokota_Horses.jpg>. Citado 3 vezes nas páginas ix, 11 e 12.
- MAUPOMÉ, D.; MEURS, M.-J. Using topic extraction on social media content for the early detection of depression. *CLEF (Working Notes)*, v. 2125, 2018. Citado na página 3.
- MIGDAL, P. *How does the $\log(p(x,y))$ normalize the point-wise mutual information?* 2015. Cross Validated. Versão da página: 24 mar. 2015. Disponível em: <<https://stats.stackexchange.com/q/143150>>. Citado na página 44.
- MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Citado 2 vezes nas páginas 12 e 13.
- MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, v. 26, p. 3111–3119, 2013. Citado 2 vezes nas páginas 8 e 11.
- MIKOLOV, T.; YIH, W.-t.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. [S.l.: s.n.], 2013. p. 746–751. Citado na página 13.
- MILLER, I. M. Rebellion, crime and violence in qing china, 1722–1911: a topic modeling approach. *Poetics*, Elsevier, v. 41, n. 6, p. 626–649, 2013. Citado na página 19.
- MISHNA, F. et al. Social media, cyber-aggression and student mental health on a university campus. *Journal of mental health*, Taylor & Francis, v. 27, n. 3, p. 222–229, 2018. Citado na página 2.
- MITCHELL, T. *Machine Learning*. [S.l.]: New York: McGraw-Hill, Inc, 1997. Citado na página 8.
- MOHR, J. W.; BOGDANOV, P. *Introduction—Topic models: What they are and why they matter*. [S.l.]: Elsevier, 2013. Citado na página 19.
- MOHR, J. W. et al. Graphing the grammar of motives in national security strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics*, Elsevier, v. 41, n. 6, p. 670–700, 2013. Citado na página 19.
- MORTENSEN, P. B. et al. Psychiatric illness and risk factors for suicide in denmark. *The Lancet*, Elsevier, v. 355, n. 9197, p. 9–12, 2000. Citado na página 1.
- NASLUND, J. et al. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, Cambridge University Press, v. 25, n. 2, p. 113–122, 2016. Citado na página 2.
- NOLES, S. W.; CASH, T. F.; WINSTEAD, B. A. Body image, physical attractiveness, and depression. *Journal of consulting and clinical psychology*, American Psychological Association, v. 53, n. 1, p. 88, 1985. Citado na página 57.

ORGANIZATION, W. H. et al. *Depression and other common mental disorders: global health estimates*. [S.l.], 2017. Citado na página 1.

OSTROWSKI, D. A. Using latent dirichlet allocation for topic modelling in twitter. In: IEEE. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. [S.l.], 2015. p. 493–497. Citado na página 19.

PATHAK, A. R.; PANDEY, M.; RAUTARAY, S. Application of deep learning for object detection. *Procedia computer science*, Elsevier, v. 132, p. 1706–1717, 2018. Citado na página 11.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 43.

PENNEBAKER, J. et al. The development and psychometric properties of liwc2007, manual. *LIWC. Austin, TX, USA*, 2007. Citado 2 vezes nas páginas 28 e 51.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 12.

PETERS, M. E. et al. Deep contextualized word representations. In: *Proceedings of NAACL-HLT*. [S.l.: s.n.], 2018. p. 2227–2237. Citado 2 vezes nas páginas 14 e 15.

PRESS, G. *A Very Short History Of Big Data*. 2013. Acesso em: 08 dez. 2020. Disponível em: <<https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/?sh=1329362565a1>>. Citado na página 18.

RADFORD, A. et al. *Improving language understanding by generative pre-training*. 2018. Citado na página 15.

RAMESH, A. et al. *DALL·E: Creating Images from Text*. 2021. Acesso em: 17 fev. 2021. Disponível em: <<https://openai.com/blog/dall-e/>>. Citado na página 11.

RAMIREZ-ESPARZA, N. et al. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In: *ICWSM*. [S.l.: s.n.], 2008. Citado 3 vezes nas páginas 3, 28 e 29.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: CITESEER. *Proceedings of the first instructional conference on machine learning*. [S.l.], 2003. v. 242, n. 1, p. 29–48. Citado na página 37.

RAUE, P. J. et al. Patients' depression treatment preferences and initiation, adherence, and outcome: a randomized primary care study. *Psychiatric Services*, Am Psychiatric Assoc, v. 60, n. 3, p. 337–343, 2009. Citado na página 1.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010. p. 45–50. <<http://is.muni.cz/publication/884893/en>>. Citado na página 36.

ŘEHŮŘEK, R.; SOJKA, P. *models.coherencemodel – Topic coherence pipeline*. 2019. Acesso em: 23 abr. 2021. Disponível em: <https://radimrehurek.com/gensim_3.8.3/models/coherencemodel.html>. Citado na página 44.

- REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. Citado 2 vezes nas páginas 18 e 41.
- REIMERS, N.; GUREVYCH, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020. Disponível em: <<https://arxiv.org/abs/2004.09813>>. Citado na página 41.
- RESNIK, P.; GARRON, A.; RESNIK, R. Using topic modeling to improve prediction of neuroticism and depression in college students. In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. [S.l.: s.n.], 2013. p. 1348–1353. Citado na página 3.
- ROCCA, J. *Understanding Variational Autoencoders (VAEs)*. 2019. Acesso em: 23 dez. 2020. Disponível em: <<https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73>>. Citado na página 24.
- RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: *Proceedings of the eighth ACM international conference on Web search and data mining*. [S.l.: s.n.], 2015. p. 399–408. Citado 2 vezes nas páginas 43 e 44.
- RONG, X. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014. Citado 2 vezes nas páginas ix e 14.
- ROSSUM, G. V.; JR, F. L. D. *Python tutorial*. [S.l.]: Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995. Citado na página 32.
- SAROKHANI, D. et al. Prevalence of depression among university students: a systematic review and meta-analysis study. *Depression research and treatment*, Hindawi, v. 2013, 2013. Citado na página 50.
- SCHNEIDER, N. et al. Chemical topic modeling: Exploring molecular data sets using a common text-mining approach. *Journal of chemical information and modeling*, ACS Publications, v. 57, n. 8, p. 1816–1831, 2017. Citado na página 19.
- SCHÖCH, C. Topic modeling genre: An exploration of french classical and enlightenment drama. *DHQ: Digital Humanities Quarterly*, v. 11, n. 2, 2017. Citado na página 19.
- SCHOMERUS, G.; MATSCHINGER, H.; ANGERMEYER, M. C. The stigma of psychiatric treatment and help-seeking intentions for depression. *European archives of psychiatry and clinical neuroscience*, Springer, v. 259, n. 5, p. 298–306, 2009. Citado na página 1.
- SERNA, M. C. et al. Duration and adherence of antidepressant treatment (2003 to 2007) based on prescription database. *European Psychiatry*, Elsevier, v. 25, n. 4, p. 206–213, 2010. Citado na página 51.
- SETH, Y. *BERT Explained – A list of Frequently Asked Questions*. 2019. Acesso em: 19 dez. 2020. Disponível em: <<https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/>>. Citado na página 16.

- SHARABI, L. L.; DELANEY, A. L.; KNOBLOCH, L. K. In their own words: How clinical depression affects romantic relationships. *Journal of Social and Personal Relationships*, Sage Publications Sage UK: London, England, v. 33, n. 4, p. 421–448, 2016. Citado na página 57.
- SILVEIRA, B.; SILVA, A. P. C. da; MURAI, F. Modelos de previsão do tom emocional de usuários em comunidades de saúde mental no reddit. In: SBC. *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*. [S.l.], 2020. p. 13–24. Citado na página 30.
- SILVER, D. et al. Mastering the game of go without human knowledge. *nature*, Nature Publishing Group, v. 550, n. 7676, p. 354–359, 2017. Citado na página 11.
- SRIVASTAVA, A.; SUTTON, C. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017. Citado 3 vezes nas páginas 23, 24 e 25.
- STEEL, Z. et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, Oxford University Press, v. 43, n. 2, p. 476–493, 2014. Citado na página 1.
- TADESSE, M. M. et al. Detection of depression-related posts in reddit social media forum. *IEEE Access*, IEEE, v. 7, p. 44883–44893, 2019. Citado 3 vezes nas páginas 3, 28 e 29.
- TANGHERLINI, T. R.; LEONARD, P. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics*, Elsevier, v. 41, n. 6, p. 725–749, 2013. Citado na página 19.
- TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 24–54, 2010. Citado na página 27.
- TOSHEVSKA, M.; STOJANOVSKA, F.; KALAJDJIESKI, J. The ability of word embeddings to capture word similarities. *International Journal on Natural Language Computing (IJNLC) Vol*, v. 9, 2020. Citado na página 12.
- TRISCOLI, C.; CROY, I.; SAILER, U. Depression predicts interpersonal problems partially through the attitude towards social touch. *Journal of affective disorders*, Elsevier, v. 246, p. 234–240, 2019. Citado na página 51.
- VANNUCCI, A.; FLANNERY, K. M.; OHANNESSIAN, C. M. Social media use and anxiety in emerging adults. *Journal of affective disorders*, Elsevier, v. 207, p. 163–166, 2017. Citado na página 1.
- VAROQUAUX, G.; GRISEL, O. Joblib: running python function as pipeline jobs. *packages.python.org/joblib*, 2009. Citado na página 43.
- VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, p. 5998–6008, 2017. Citado na página 15.
- VINYALS, O. et al. *AlphaStar: Mastering the Real-Time Strategy Game StarCraft II*. 2019. <<https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/>>. Citado na página 11.

- VOGEL, E. A. et al. Social comparison, social media, and self-esteem. *Psychology of Popular Media Culture*, Educational Publishing Foundation, v. 3, n. 4, p. 206, 2014. Citado na página 1.
- WANG, X.; GRIMSON, E. Spatial latent dirichlet allocation. *Advances in neural information processing systems*, v. 20, p. 1577–1584, 2007. Citado na página 19.
- WANG, Y. et al. Family dysfunction and adolescents' anxiety and depression: A multiple mediation model. *Journal of Applied Developmental Psychology*, Elsevier, v. 66, p. 101090, 2020. Citado 2 vezes nas páginas 49 e 63.
- WELLING, M. Product of experts. *Scholarpedia*, v. 2, n. 10, p. 3879, 2007. Revision #137078. Citado na página 25.
- WHITE, H. *Artificial neural networks: approximation and learning theory*. [S.l.]: Blackwell Publishers, Inc., 1992. Citado na página 11.
- WOODS, H. C.; SCOTT, H. # sleepteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of adolescence*, Elsevier, v. 51, p. 41–49, 2016. Citado na página 1.
- WORSHAM, J.; KALITA, J. Genre identification and the compositional effect of genre in literature. In: *Proceedings of the 27th International Conference on Computational Linguistics*. [S.l.: s.n.], 2018. p. 1963–1973. Citado na página 11.
- YAMADA, I. et al. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. [S.l.]: Association for Computational Linguistics, 2020. p. 23–30. Citado na página 41.
- YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: NASHVILLE, TN, USA. *Icml*. [S.l.], 1997. v. 97, n. 412-420, p. 35. Citado na página 37.
- ZÚÑIGA, H. Gil de; MOLYNEUX, L.; ZHENG, P. Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. *Journal of communication*, Oxford University Press, v. 64, n. 4, p. 612–634, 2014. Citado na página 1.

Apêndices