UFRRJ

INSTITUTO MULTIDISCIPLINAR

PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

DISSERTAÇÃO

Classificação de Publicações em Humanidades Digitais Apoiada em Abordagem Taxonômica

Luiz Carlos de Jesus

2022

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

CLASSIFICAÇÃO DE PUBLICAÇÕES EM HUMANIDADES DIGITAIS APOIADA EM ABORDAGEM TAXONÔMICA

LUIZ CARLOS DE JESUS

Sob orientação de
Ricardo Cordeiro Corrêa

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências** no Programa de Pós-graduação Interdisciplinar em Humanidades Digitais, Área de Concentração em Análise Qualitativa e Quantitativa de Dinâmicas Sociais.

Nova Iguaçu, RJ
Abril de 2022

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

J58c

Jesus, Luiz Carlos de, 1971-
Classificação de Publicações em Humanidades Digitais Apoiada em Abordagem Taxonômica / Luiz Carlos de Jesus. - Nova Iguaçu, 2022.
90 f.: il.

Orientador: Ricardo Cordeiro Corrêa.
Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais, 2022.

1. Humanidades Digitais. 2. Processamento de Linguagem Natural. 3. Classificação de Textos. 4. Taxonomia. I. Corrêa, Ricardo Cordeiro, 1966-, orient. II Universidade Federal Rural do Rio de Janeiro. Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais III. Título.

Este documento foi criado usando o sistema LaTeX de preparação de documentos desenvolvido por Leslie Lamport a partir do sistema de formatação TeX criado por Donald Knuth.
O formato foi obtido usando a classe `UFRuralRJ`, uma adaptação livre das classes `mdtufsm` e `iiufrgs` para a formatação de documentos acadêmicos produzidos na Universidade Federal Rural do Rio de Janeiro (UFRRJ).

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

LUIZ CARLOS DE JESUS

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências** no Programa de Pós-graduação Interdisciplinar em Humanidades Digitais, Área de Concentração em Análise Qualitativa e Quantitativa de Dinâmicas Sociais.

DISSERTAÇÃO APROVADA EM 28/04/2022.

Ricardo Cordeiro Corrêa. Docteur UFRRJ (Orientador)
Alexandre Fortes. D.Sc. UFRRJ
Carlos Eduardo Ribeiro de Mello. D.Sc. UNIRIO
Lucas Correia Carvalho. D.Sc. UFF

ATA DE DEFESA DE TESE Nº 152/2022 - PPGIHD (11.39.00.16)

Nº do Protocolo: 23083.032446/2022-40

Nova Iguaçu-RJ, 27 de maio de 2022.

Visualize o documento original em https://sipac.ufrrj.br/public/documentos/index.jsp informando seu número: **152**, ano: **2022**, tipo: **ATA DE DEFESA DE TESE**, data de emissão: **27/05/2022** e o código de verificação: **10ae3adf50**

(Assinado digitalmente em 27/05/2022 09:35)
ALEXANDRE FORTES
PROFESSOR DO MAGISTERIO SUPERIOR
DeptH/IM (12.28.01.00.00.88)
Matrícula: ###084#6

(Assinado digitalmente em 27/05/2022 05:51)
RICARDO CORDEIRO CORREA
PROFESSOR DO MAGISTERIO SUPERIOR
PPGIHD (11.39.00.16)
Matrícula: ###07#4

(Assinado digitalmente em 23/08/2022 17:20)
CARLOS EDUARDO RIBEIRO DE MELLO
ASSINANTE EXTERNO
CPF: ###.###.927-##

(Assinado digitalmente em 31/05/2022 09:05)
LUCAS CORREIA CARVALHO
ASSINANTE EXTERNO
CPF: ###.###.237-##

Emitido em 05/05/2023

TERMO Nº 489/2023 - PPGIHD (11.39.00.16)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 08/05/2023 00:52)
RICARDO CORDEIRO CORREA
COORDENADOR CURS/POS-GRADUACAO - TITULAR
PPGIHD (11.39.00.16)
Matrícula: ###07#4

Visualize o documento original em https://sipac.ufrrj.br/documentos/ informando seu número: **489**, ano: **2023**, tipo: **TERMO**, data de emissão: **08/05/2023** e o código de verificação: **8fde279a42**

AGRADECIMENTOS

Agradeço primeiramente, a Deus por todas as bênçãos recebidas. A minha esposa Sônia e ao meu filho Abner pelo apoio, a paciência e os cuidados comigo em mais essa jornada.

Ao meu professor e orientador Ricardo Corrêa, o meu agradecimento especial, por aceitar orientar-me no desbravamento deste trabalho, e mais ainda, por sua dedicação e a austeridade que se traduz como um exemplo que desejo seguir em minha trajetoria acadêmica e na vida.

Aos amigos, irmão e incentivadores que sustentam a minha sanidade nos momentos mais difíceis, me auxiliando e ajudando para que assim eu possa ter as minhas força recarregadas e retomar a trajetória para seguir adiante.

A Universidade Federal Rural do Rio de Janeiro, onde eu me sinto estando em meu segundo lar, ambiente agradável e afetuoso, repleto de boas recordações, de mestres zelosos e de muitos trabalhos dedicados na manutenção do ensino de excelência. Aos professores do Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais, que constrói esse sonho diariamente, minha eterna gratidão por esta enorme oportunidade e seus esforço em desenvolver pessoas na Baixada Fluminense.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO GERAL

DE JESUS, Luiz Carlos. **Classificação de Publicações em Humanidades Digitais Apoiada em Abordagem Taxonômica**. 2022. 90f. Dissertação (Mestrado em Humanidades Digitais). Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro, Nova Iguaçu, RJ, 2022.

As publicações em periódicos científicos e em conferências especializadas desempenham o papel primordial de expressar os temas de interesse de autores e leitores em um determinado campo do conhecimento. Nesse sentido, o esforço de organizar a produção científica é vital para o avanço da difusão dos conteúdos produzidos de forma inequívoca, rápida e segura. Considerando a atual inundação informacional provocada pelas ferramentas digitais, a questão da classificação automatizada se torna premente e deve obrigatoriamente ser abordada em todo repositório ou plataforma digital de publicações científicas. Dentre outros aspectos, sobressai-se o uso de uma taxonomia pela sua capacidade de adicionar um elemento semântico hierárquico ao ato de classificar ou categorizar conceitos e informações específicas que definem o domínio de um campo do conhecimento. Particularmente no campo das Humanidades Digitais, a cultura epistemológica que vem sendo construída pela sua crescente comunidade tem feito nascer e crescer projetos internacionais que abordam a questão em um ambiente com desafios adicionais devido ao seu perfil fortemente interdisciplinar. O objetivo desta dissertação é usar ferramentas computacionais de análise por tópicos de textos para desenvolver um método auxiliar de classificação léxica de publicações apoiado em uma taxonomia denominada *TaDiRAH – Taxonomy of Digital Research Activities in the Humanities*. O método proposto pode ser visto como uma combinação da abordagem semântica da taxonomia com a abordagem léxica da análise automatizada de textos. Suas categorias são de uso livre e prático. No entanto, não é incomum, e até esperado pelo perfil interdisciplinar, que uma publicação possa ser classificada em diferentes categorias de níveis diferentes ou de mesmo nível da taxonomia, criando assim sobreposições. Somado a isso, a quantidade de publicações já classificadas artesanalmente pela comunidade científica ainda é relativamente pequena e, sobretudo, extremamente desbalanceada entre as categorias da taxonomia. Esses dois aspectos que caracterizam a amostragem disponível tornam a tarefa de classificar com fidedignidade publicações em Humanidades Digitais particularmente difícil. Propomos um método que combina modelos de classificação bayesianos da literatura com abordagens originais para lidar com sobreposições e desbalanceamento entre as categorias da taxonomia. Resultados de experimentos computacionais realizados com um universo de 443 publicações mostraram que as abordagens propostas são, de fato, capazes de melhorar profundamente o desempenho dos métodos de classificação empregados.

Palavras-chave: Humanidades Digitais, Processamento de Linguagem Natural, Classificação de Textos, Taxonomia.

GENERAL ABSTRACT

DE JESUS, Luiz Carlos. **Taxonomic-Based Digital Humanities Publications Classification**. 2022. 90p. Dissertation (Master in Digital Humanities). Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro, Nova Iguaçu, RJ, 2022.

Publications in scientific journals and in specialized conferences play a key role in expressing the topics of interest to authors and readers in a given field of knowledge. In this sense, the effort to organize scientific production is vital for the advancement of the dissemination of contents produced in an unequivocal, fast and safe way. Considering the current information flood caused by digital tools, the issue of automated classification becomes urgent and must be addressed in every repository or digital platform of scientific publications. Among other aspects, the use of a taxonomy stands out for its ability to add a hierarchical semantic element to the act of classifying or categorizing concepts and specific information that define the domain of a field of knowledge. Particularly in the field of Digital Humanities, the epistemological culture that has been built by its growing community has given rise to international projects that address the issue in an environment with additional challenges due to its strongly interdisciplinary profile. The objective of this dissertation is to use computational tools for analysis by topics of texts to develop an auxiliary method of lexical classification of publications supported by a taxonomy called *TaDiRAH – Taxonomy of Digital Research Activities in the Humanities*. The proposed method can be seen as a combination of the semantic approach of taxonomy with the lexical approach of automated text analysis. Its categories are free and practical. However, it is not uncommon, and even expected by the interdisciplinary profile, that a publication can be classified into different categories of different levels or the same level of the taxonomy, thus creating overlaps. In addition, the number of publications already classified by the scientific community is still relatively small and, above all, extremely unbalanced between the taxonomy categories. These two aspects that characterize the available sample make the task of reliably classifying publications in Digital Humanities particularly difficult. We propose a method that combines Bayesian classification models from the literature with original approaches to deal with overlaps and imbalances between taxonomy categories. Results of computational experiments carried out with a universe of 443 publications showed that the proposed approaches are, in fact, capable of profoundly improving the performance of the classification methods use.

Keywords: Digital Humanities, Natural Language Processing, Text Classification, Taxonomy.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Humanidades Digitais são vistas, em seu sentido amplo, como o terreno comum entre a Ciência da Computação, a Tecnologia da Informação e Comunicação e os problemas e temas de pesquisa em Ciências Humanas. Desde a década de 90 do século XX, a cultura epistemológica para esse campo vem sendo construída em torno do princípio segundo o qual métodos e tecnologias computacionais são usados na modelagem de fenômenos sociais e da compreensão que deles temos (UNSWORTH, 2013). Os elementos essenciais nesse casamento são dados oriundos das diversas formas de representação, em meios digitais variados, dos registros deixados pelos fenômenos estudados e suas associações, com forte engajamento no aspecto "registros digitais" como objeto de estudo. Dessa forma, o terreno comum das Humanidades Digitais é preenchido com uma vasta gama de ferramentas e abordagens metodológicas para lidar com dados, representações, múltiplas mídias e modelos. Diante desses aspectos, há autores que descrevem o campo das Humanidades Digitais como um conjunto de práticas interdisciplinares nas quais computadores e algoritmos são usados como ferramenta. No entanto, o seu amadurecimento vem tornando progressivamente nítido um cenário de posições teóricas próprias no qual as metodologias computacionais de modelagem e análise de dados de humanidades fornecem elementos que, potencialmente, expandem a capacidade de compreensão humana, individual ou coletiva, de fenômenos sociais. Essas são qualidades tipicamente associadas ao estabelecimento de um novo campo de conhecimento, evidenciado também pela existência de departamentos ou unidades acadêmicas, organizações profissionais, conferências regulares, periódicos científicos, revistas, livros e programas de incentivo a atividades colaborativas (TERRAS; NYHAN; VANHOUTTE, 2013).

Em sua relativamente breve história, o campo das Humanidades Digitais segue a tendência de constantemente se auto observar, criando mecanismos, métodos e técnicas de avaliar, sistematizar e indexar as suas concepções e atividades. Papel primordial nesse movimento desempenham as publicações em periódicos científicos e em conferências especializadas ao expressar os temas de interesse dos autores e leitores envolvidos nesse campo do conhecimento. Os meios de divulgação científica foram criados com o objetivo de promover um diálogo orgânico entre os diferentes parceiros interessados nesses conhecimentos e divulgar as publicações que registram esse conhecimento disseminando-as através dos periódicos científicos, reconhecidos propagadores, cujo contexto moderno relaciona-se também com as bases de dados digitais (BOREK et al., 2014; MACHADO, 2016; VITAL; CAFÉ, 2011). Nesse sentido, as publicações são os registros digitais que servem de base comparativa para descrever e analisar o perfil dos temas, objetos, métodos e atividades que expressam os campos de abordagem das Humanidades Digitais em diferentes recortes de tempo e espaço (GRIFFITHS; STEYVERS, 2004).

O esforço de organizar a produção científica é vital na sua análise e posterior seleção e recuperação dos conteúdos de forma inequívoca, rápida e segura para uso apropriado nos projetos de Humanidades Digitais (BOREK et al., 2014; CASTRO, 2020). Nesse contexto, o uso de uma taxonomia adiciona um elemento semântico hierárquico à análise hermenêutica de textos que possibilita classificar ou categorizar as publicações segundo conceitos e informações específicas (BOREK et al., 2016). As taxonomias são as ferramentas indicadas quando se trata de organizar informações e conhecimentos em ambientes

digitais com o objetivo de buscar consistência e uniformidade na publicação, na recuperação e no consenso terminológico. Além disso, segundo (BOREK et al., 2016), uma taxonomia pode ser vista como um ponto de partida para o desenvolvimento de uma ontologia (VITAL; CAFÉ, 2011), em uma abordagem que transita do mais pragmático para o mais erudito. Trata-se, portanto, de uma analogia conceitual com o papel que as taxonomias desempenham, por exemplo, no entendimento da biodiversidade em diferentes níveis, identificando, quantificando e catalogando os seres vivos estudados de forma comparativa entre si (CHAMORRO et al., 2018). A sua inclusão no processo de classificação, dada a inundação informacional provocada pelas ferramentas digitais tão presentes em nossa sociedade, aponta para a integração entre as tecnologias informacionais e a análise textual no âmbito da antiga polarização, constantemente atualizada, entre abordagens quantitativas e qualitativas.

Os métodos de análise automatizada de textos atuam sobre propriedades léxicas e estruturais do que semânticas. Esse fato é uma decorrência de que essas propriedades são mais diretamente expressas em estruturas matemáticas próprias à concepção de algoritmos. Assim sendo, a principal questão associada à integração de tecnologias informacionais à análise hermenêutica está na busca por resultados algorítmicos de caráter explicativo que forneçam elementos complementares à compreensão humana das informações contidas em grandes quantidades de documentos. No caso específico da análise de publicações científicas, trata-se de construir ferramentas computacionais que produzam resultados úteis e eficazes suplementares que, conjugados com a experiência acumulada de análise hermenêutica feita por esepecialistas, produzam resultados de classificação com alto grau de precisão, amplitude e eficiência, e com custos humanos e de tempo reduzidos. A busca por atingir esse objetivo não segue, necessariamente, o caminho da reprodução da capacidade humana de análise hermenêutica, abordagem essa que possui obstáculos que podem ser intransponíveis. A abordagem alternativa a seguir seria a de buscar indicadores que possam ser calculados automaticamente e que sejam informativos, baseados em propriedades do texto que não sejam derivadas diretamente da semântica presente nos textos, mas que produzem um efeito semântico no conteúdo dos documentos (GRIFFITHS; STEYVERS, 2004).

Considerando o cenário traçado acima, um problema naturalmente formulado é o do estabelecimento de um procedimento computacional de classificação de publicações em Humanidades Digitais segundo os estratos definidos em uma taxonomia desse campo do conhecimento. Dentro desse amplo tema, o escopo da contribuição desta dissertação é definido pelo conjunto de publicações analisado, pela taxonomia adotada para guiar a classificação e pela abordagem de análise textual empregada. O envolvimento das plataformas de divulgação científicas é um elemento primordial para tornar a classificação em âmbito mundial, já que temos percebido que existe um esforço considerável das bases e portais em disponibilizar cada vez mais elementos com a expressão de busca “Digital Humanities”. Porém, o envolvimento de um conjunto de especialistas da área continua a ser essencial para classificar com fidedignidade publicações nesse campo. Por isso, a importância de termos uma ferramenta auxiliar que nos possibilite fazer essa análise com a amplitude mundial.

O acesso digital a publicações científicas originadas no Brasil e no exterior, com os mais variados perfis, é atualmente possível através de plataformas concebidas para esse fim (MACHADO, 2016). Dentre elas, pode-se citar aquelas de propósito genérico de acordo com a origem (mantidos por universidades, órgãos públicos e organizações com ou sem

fins lucrativos) ou dedicadas a algum domínio de especialidades específico, tais como o *PubMed*[1] (Medicina), *ArXiv*[2] (física, matemática, computação, estatística e biologia), *BPubs*[3] (negócios e mercado), *SciELO*[4] (mais voltados para América Latina e do Caribe), *Scopus*[5] (multidisciplinar) e *Web of Science*[6] (multidisciplinar), dentre outras.

O conjunto de amostras utilizado nas análises desta dissertação é composto por 443 publicações redigidas em inglês, classificadas artesanalmente por especialistas em Humanidades Digitais e coletadas no grupo *Doing Digital Humanities – a DARIAH bibliography* plataforma *Zotero*[7]. Esse grupo sintetiza o resultado de uma atividade contínua do projeto *DARIAH-EU*[8] que consiste em coletar, referenciar e classificar publicações sobre questões de Humanidades Digitais. Esse projeto é organizado em rede, vinculada à União Europeia, constituída por especialistas que reúnem informação, conhecimento, conteúdo, métodos, ferramentas e tecnologias de apoio à pesquisa e ao ensino em Artes e Humanidades fazendo uso de recursos digitais. Ao formar a base dos experimentos deste trabalho, o mesmo pode também ser visto como uma ação de emprego da taxonomia a fim de fornecer informações catalogadas de maneira automatizada sobre a produção científica em Humanidades Digitais e ao mesmo tempo engajar pessoas com os objetivos dessa disciplina (BOREK et al., 2016; DARIAH-UE CONSORTIUM, 2012).

Do mesmo projeto *DARIAH-EU* é a taxonomia usada, denominada *TaDiRAH - Taxonomy of Digital Research Activities in the Humanities*[9]. Suas estruturação hierárquica e produção vêm ocorrendo de forma totalmente colaborativa e expressam a síntese do campo das Humanidades Digitais que, além do objetivo de respaldar a tarefa de organizar, catalogar e resgatar informações relevantes e essenciais ao avanço das pesquisas, é também uma potente ferramenta de oportunidades para o desenvolvimento, a integração e a colaboração interdisciplinar no campo, colaborando para o fortalecimento dos fazeres e dos saberes dos usuários envolvidos (BOREK et al., 2014, 2016; CASTRO, 2020; TADIRAH, 2014; VIGNOLI; SOUTO; CERVANTES, 2013). Seus estratos são de uso livre e prático, podendo um elemento ser classificado em diferentes estratos de níveis diferentes na hierarquia ou de mesmo nível (BOREK et al., 2020). Devido ao grau de participação da comunidade, a *TaDiRAH* é altamente especializada para as Humanidades Digitais e foi elaborada para atender a necessidade prioritária do uso prático facilitando a sua adoção em diferentes contextos (BOREK et al., 2016).

A existência e a relevância de uma taxonomia, como a *TaDiRAH*, só se justifica se efetivamente empregada para classificar os registros da produção científica através de publicações especializadas na forma de artigos e livros (MACHADO, 2016). A relativa pequena quantidade de publicações atualmente classificadas segundo a *TaDiRAH* revela a dificuldade prática decorrente da necessidade de envolver um conjunto expressivo de especialistas. Nesse contexto, uma ferramenta que proporcione algum grau de automatismo à tarefa torna o processo viável. A resposta fornecida pelo método proposto pode vir a servir de

[1] <https://pubmed.ncbi.nlm.nih.gov/>
[2] <https://arxiv.org/>
[3] <http://www.bpubs.com/>
[4] <https://www.scielo.org/>
[5] <https://www.scopus.com/home.uri>
[6] <https://www.webofscience.com>
[7] <https://www.zotero.org/groups/113737/doing_digital_humanities_-_a_dariah_bibliography/collections/M2GQV3T4>
[8] <https://www.dariah.eu/>
[9] <https://vocabs.dariah.eu/tadirah/en/>

base para a classificação definitiva por meio da intervenção posterior de especialistas da comunidade científica.

O objetivo desta dissertação é, em síntese, propor e experimentar ferramentas computacionais de análise textual (BLEI; MCAULIFFE, 2007; BLEI; NG; JORDAN, 2003; BLEI; LAFFERTY, 2006b; ROBERTS et al., 2013) para desenvolver um método auxiliar de classificação léxica de publicações apoiado em uma taxonomia que define uma estrutura hierárquica não disjunta de estratos. Nossa abordagem consiste em classificar publicações científicas de Humanidades Digitais a partir de seus títulos e resumos. O método proposto pode ser visto como uma combinação da abordagem semântica da taxonomia com a abordagem léxica da análise automatizada de textos (BOREK et al., 2016; GRIFFITHS; STEYVERS, 2004; WELBERS; ATTEVELDT; BENOIT, 2017). Para atingir esse objetivo, seguimos duas abordagens em associação: a proposta de um método de classificação bayesiano com ferramentas computacionais de análise por tópicos cuja função é descobrir padrões de uso das palavras em estruturas textuais (BLEI; NG; JORDAN, 2003; ROBERTS et al., 2013), e uma taxonomia, a *TaDiRAH*, para fornecer a base de estratos descrevendo uma organização da produção científica em Humanidades Digitais. Nossa finalidade é registrar o conhecimento dentro desse domínio entendendo que não existe uma relação biunívoca entre as publicações a serem classificadas e os estratos da taxonomia *TaDiRAH*. Isto acontece porque uma publicação pode ser classificada em diferentes estratos e em diferentes níveis ao mesmo tempo, ou seja, a *TaDiRAH* é uma estrutura hierárquica não disjunta de classificação (BOREK et al., 2016).

Para realizar a análise das publicações escolhemos ferramentas de análise por tópicos que é uma metodologia estatística que busca expressar a estrutura de um corpus através da leitura distante (TEMPLETON, 2011), conhecidas como modelos generativos probabilísticos (LIU et al., 2016). A análise por tópicos tenta imitar o que é o processo de escrita humana (ALGHAMDI; ALFALQI, 2015) e é usada para melhorar a nossa compreensão sobre os movimentos temáticos de um texto (GRIFFITHS; STEYVERS, 2004), distinguindo as suas estruturas ocultas (YAU et al., 2014). Dentre os diversos modelos existentes na literatura, adotamos o modelo *Latent Dirichlet Allocation – LDA* e seus descendentes pelas seguintes razões: (1) É mais citado e usado, inclusive com outras abordagens (BLEI, 2012); (2) Serve de referência para outros modelos (BLEI; LAFFERTY, 2007; SHEN; SUN; SHEN, 2008); (3) Possui diversidade de trabalhos para análise e revisão (GRIFFITHS; STEYVERS, 2004); (4) É ajustável a diversos tipos de corpus (MCCALLUM; WANG; CORRADA-EMMANUEL, 2007); (5) É adaptável a diversos contextos (VAYANSKY; KUMAR, 2020); (6) Possui diversas versões e serve de base para outros modelos (ALGHAMDI; ALFALQI, 2015), inclusive para classificação de textos (BLEI; LAFFERTY, 2007; BLEI; MCAULIFFE, 2007; GRIFFITHS; STEYVERS, 2004; ROBERTS et al., 2013).

O interesse crescente dos pesquisadores em aplicar modelos estatísticos hierárquicos baseados em tópicos, inclusive para classificação, se explica pela crescente complexidade dos corpora de análise particularmente em grandes coleções de documentos disponibilizadas em formato eletrônico (BLEI; LAFFERTY, 2009). No caso específico do corpus de publicações em Humanidades Digitais a partir da *TaDiRAH*, há desafios adicionais em relação à classificação de publicações de outras temáticas ou quando baseados em outras estruturas de classes. Esse fato é decorrência do desafiador percurso da classificação automatizada em face (i) das características interdisciplinares do campo, refletida nos trabalhos que podem ser classificados em diversos estratos diferentes da taxonomia *TaDiRAH* (BOREK

et al., 2016) e (ii) da relativamente reduzida e desbalanceada amostra de publicações já classificadas. Para lidar com essa da complexidade do corpus, empregamos dois métodos derivados da família do LDA que além de oferecer as estatísticas descritivas úteis que facilitam as análises, também pode classificar grandes quantidades de documentos similares (BLEI; MCAULIFFE, 2007; ROBERTS et al., 2013). Tendo em vista que os modelos são adaptáveis a qualquer coleção de documentos, incluindo os nascidos digitalmente (GRIFFITHS; STEYVERS, 2004) e pode ser usado para classificar textos em diferentes versões e contextos (BAI; WANG, 2015; HINGMIRE et al., 2013), tal como em Cienciometria, por exemplo (YAU et al., 2014).

A maior contribuição deste trabalho é uma metodologia de adaptação, configuração e combinação de métodos computacionais conhecidos com o intuito de lidar com as especificidades da aplicação escolhida. O conceito computacional central na abordagem seguida é o de análise por tópicos, que consiste em conjuntos de técnicas que tem por finalidade encontrar padrões estruturais e temas relacionados às frequências das palavras dentro de um texto (VAYANSKY; KUMAR, 2020). Dentre os diversos métodos conhecidos para análise de tópicos adotamos parcimoniosamente métodos da família do modelo *Latent Dirichlet Allocation*. Os principais desafios enfretados são divididos em duas etapas. Em primeiro lugar, a existência de estratos com sobreposição devido a publicações classificadas em mais de um estrato é um fato raramente encontrado nas aplicações dos métodos de classificação textual. Nesse sentido, propomos um método para lidar com tal cenário adaptado ao caso específico da classificação baseada na *TaDiRAH*. O segundo desafio decorre da propriedade de estratos desbalanceados - com quantidades díspares de publicações por estrato, chegando à proporção de em torno de 9 vezes de um estrato para o outro. O método proposto para contornar esse problema é totalmente inovador no princípio de gerar amostras sintéticas dos estratos minoritários utilizando o modelo generativo da abordagem *LDA* original. Adicionalmente a esses desafios metodológicos, o conjunto de publicações já classificadas restrito e relativamente pequeno – dificultando o aprendizado de máquina – é um desafio adicional quanto ao procedimento de validação empírica das propostas realizadas. Dentro desses limites, os métodos apresentam valores de medidas adequadas de eficácia nos experimentos realizados.

Antes de classificar, propriamente dito, testamos diversos outros métodos dos modelos de aprendizado de máquina não supervisionados, semisupervisionados e supervisionados (ALGHAMDI; ALFALQI, 2015; GRIFFITHS; STEYVERS, 2004; VAYANSKY; KUMAR, 2020; WATANABE; ZHOU, 2020), e chegamos a conclusão de que para a nossa pesquisa a abordagem mais adequada realmente seria, a opção por modelos da família do LDA. Com esses modelos em mãos, classificamos os textos com os rótulos que consideramos mais apropriados, também pudemos produzir textos sintéticos (BATISTA; PRATI; MONARD, 2004), usando a abordagem de tópicos para criar textos com os mesmo perfis e padrões dos textos originais. Essas amostras sintéticas foram necessárias para aumentar o conjunto minoritário de textos para que o modelo pudesse aprender melhor as suas diferenças em relação ao conjunto de textos da classe majoritária, e assim fazer distinção entre eles para poder identificá-los de forma mais proveitosa para a nossa pesquisa.

Desejamos neste trabalho, estimular a reflexão do uso associativo de abordagens semânticas (taxonomia) e análise léxica por tópicos, usando a *TaDiRAH*, no contexto das Humanidades Digitais, para o entendimento e o desenvolvimento do campo (CASTRO, 2020), assim como já temos exemplos de uso de ontologias com processamento de linguagem natural em (LUZ, 2013). Os nossos resultados apontam para vantagens na adoção definitiva

da classificação automatizada de textos, dado que já temos bases de distribuição de publicações científicas que já a usam, porém não em associação com a *TaDiRAH*. Se a adoção do método vier a ser uma realidade na classificação automatizada, para a comunidade de humanistas digitais, isso terá retorno direto na velocidade das pesquisas e no retorno do pesquisador em classificar os seus trabalhos de acordo com a taxonomia, o que significa que poderemos ter ainda mais trabalhos já produzidos com a classificação segundo a *TaDiRAH*, tendo mais publicações para treino e teste do algoritmo e melhorando cada vez mais a classificação automatizada. Isso também diminuiria os problemas causados pelo desbalanceamento das classes, melhorando ainda mais a eficácia do método. Por fim, esperamos estimular o uso dos estratos da *TaDiRAH* pela comunidade a fim de termos mais textos disponíveis para a revisão *ad hoc* posterior por especialistas e trabalhos futuros.

Este trabalho está organizado da seguinte forma. O tema tratado no Capítulo 2 está relacionado à taxonomia *TaDiRAH*. A abordagem seguida consiste em apresentar os conceitos gerais de sistemas de organização do conhecimento. Essa apresentação inicial permite estabelecer um direcionamento que apoie o uso da *TaDiRAH* como uma ferramenta útil e importante para o campo das Humanidades Digitais. O capítulo é concluído com uma análise da estrutura e dos textos descritivos da taxonomia, seguida de uma análise prévia dos textos previamente classificados recolhidos no contexto das Humanidades Digitais com o objetivo de embasar a abordagem adotada.

No Capítulo 3, abordamos o uso da análise automatizada de textos. A abordagem consiste em apresentar as ferramentas que usamos para classificar as publicações no âmbito da abordagem de tópicos. Essa apresentação inicial permite estabelecer as bases para o entendimento do leitor a respeito dos tipos de modelos existentes e da opção feita para o nosso trabalho.

O Capítulo 4 está relacionado ao conjunto de dados da análise. A abordagem seguida consiste em, primeiramente, apresentar o caminho que percorremos para realizar o experimento. Essa apresentação permite estabelecer uma visão nítida de todo o experimento, das dificuldades encontradas e das soluções tomadas para a sua realização no contexto das sobreposições de estratos. Usamos um modelo de classificação bayesiana que associa-se a outros de aprendizado de máquina (BLEI; MCAULIFFE, 2007; BLEI; NG; JORDAN, 2003; BLEI; LAFFERTY, 2006b; ROBERTS et al., 2013) que associa os temas ocultos dos metadados das publicações aos estratos da *TaDiRAH*. O capítulo é concluído com uma análise sintética do experimento.

O Capítulo 5 trata da questão do desbalanceamento entre os estratos e do mecanismo adotado para superar essa dificuldade. Terminamos o capítulo demonstrando como conseguimos melhorar ainda mais os resultados dos experimentos. Todos os experimentos foram levados a cabo com codificações dos métodos apresentados na linguagem R efetuadas pelo professor orientador.

No Capítulo 6, apresentamos as nossas conclusões tendo como base uma síntese geral das conclusões de todo processo e análise final. um rápido resumo seguido das conclusões de nossos experimentos. Também apontamos as oportunidades de melhorias que visualizamos em trabalhos futuros.

2 TAXONOMIA DAS ATIVIDADES EM HUMANIDADES DIGITAIS

Neste capítulo, tratamos da taxonomia de Humanidades Digitais usada concomitantemente a ferramentas computacionais de análise automatizada de textos para classificação de publicações científicas. Começamos com uma visão geral contextual da estrutura de estratos e utilização da *TaDiRAH*, ressaltando de forma sucinta a sua importância e relevância no contexto das Humanidades Digitais. Em seguida, em uma análise prévia, são apresentadas características gerais relevantes ao uso da *TaDiRAH* para classificação automática de textos com duas abordagens, a saber: (1) uma abordagem quantitativa com a qual mostramos a estrutura da frequência de ocorrência nas publicações classificadas e dos textos descritivos dos estratos da taxonomia; e (2) uma abordagem qualitativa, mostrando esse comportamento como se dá no âmbito das técnicas multidisciplinares próprias do campo das Humanidades Digitais. O capítulo é encerrado com uma síntese geral da estrutura da taxonomia, os comportamentos mais esperados e uma visão sistêmica dos desafios e das possíveis soluções que podemos adotar no percurso deste trabalho.

Ao longo do capítulo, e de resto nos demais capítulos da dissertação, denominamos de um *elemento léxico* uma expressão simples ou composta envolvendo uma ou mais palavras dotada de um significado específico. A título de ilustração, expressões compostas como *digital humanities*, *content analysis* e *data cleansing* são exemplos de elementos léxicos em inglês, assim como *transcribing*, *recording* e *imaging* são exemplos formados por expressões simples. Outros exemplos podem ser encontrados no Capítulo 4, ocasião na qual são mencionados algoritmos de determinação de elementos léxicos a partir da detecção de ocorrências subsequentes frequentes de palavras (MANNING; SCHÜTZE, 1999).

2.1 Motivações e Objetivos da *TaDiRAH*

Ademais de ser admitida como um veículo de aproximação entre os centros de humanidades e, ao mesmo tempo, como uma forma de compreensão do campo de Humanidades Digitais, a *TaDiRAH* também é usada como uma ferramenta para o desenvolvimento do seu próprio domínio. Segundo (CASTRO, 2020), a *TaDiRAH* pode ser vista como a estrutura da prática e do conceito do campo das Humanidades Digitais definidos (i) nas atividades ou ações práticas de pesquisa, tais como digitalizar, comentar ou analizar documentos; (ii) nas técnicas ou procedimentos de pesquisa pró resultados mediados pelas tecnologias, tais como usar um código pronto ou produzir um algoritmo próprio; (iii) e nos objetos digitais de pesquisa, tais como documentos, imagens e informações nascidos digitalmente ou digitalizados (PERKINS et al., 2014).

A busca pela compreensão do perfil e da dinâmica do campo das Humanidades Digitais tem na *TaDiRAH* uma ferramenta útil no sentido em que ela venha a ser empregada para classificar e categorizar as atividades que compõem esse campo (CASTRO, 2020). No cerne do que é a *TaDiRAH* vêm embutidos seus objetivos e seu uso. Graças à sua objetividade e praticidade podemos entender melhor como a taxonomia ajuda a desenvolver o campo das Humanidades Digitais, pelo seu uso eletrônico, pela coleta e estruturação feita nas informações mais relevantes a fim de torná-las mais facilmente detectáveis, leve, colaborativa, objetiva e expansível (BOREK et al., 2014). Desde o seu lançamento, ela vem sendo adotada e usada pela comunidade acadêmica do campo em diferentes contextos,

desenvolvendo diferentes trabalhos.

Podemos refletir sobre a importância da *TaDiRAH* para o campo das Humanidades Digitais a partir da resposta imediata da comunidade ante ao lançamento da taxonomia. No repositório do projeto na plataforma *github* (DHTAXONOMY, 2020), podemos notar a expectativa dos criadores da *TaDiRAH* diante das 8 iniciativas então mapeadas por eles em sua página. Essa expectativa da comunidade das Humanidades Digitais tem consistência na sua necessidade de integração para facilitar o acesso às informações a respeito dos pesquisadores e suas atividades favorecendo o desenvolvimento do campo. Um exemplo disso pode ser notado na forma em que os desenvolvedores da taxonomia adotaram para garantir feedbacks através de rascunhos públicos (BOREK et al., 2016), uma viabilização prática do Manifesto das Humanidades Digitais expressa em (DACOS, 2011) nos itens 5 e 6, e demonstra uma tendência do campo. Na versão 2.0 a *TaDiRAH* possui ramificações em alemão, francês, espanhol, português, sérvio e italiano. Comparativamente à taxa de crescimento do campo, podemos considerar como baixa a adesão dos pesquisadores ao uso da *TaDiRAH*, porém seus impactos podem ser sentidos através dos mecanismos de debate e análise do desenvolvimento do campo, isso porque a *TaDiRAH* tem se tornado um eficiente mecanismo de entendimento e análise epistemológica das Humanidades Digitais, tanto para os de dentro quanto para os de fora do campo.

2.2 Estratos da *TaDiRAH*

A *TaDiRAH* é constituída em seu primeiro nível por 7 estratos mais abrangentes (BOREK et al., 2021; TADIRAH, 2014), a saber:

- *Analyzing* (Figura 2.1): Refere-se à extração de qualquer tipo de informação a partir dos dados, elementos, padrões, agrupamentos e descoberta de fenômenos informacionais pontuais ou recorrentes, seja em aspectos estruturais, visualização de resultados, formais ou semânticos.
- *Capturing* (Figura 2.1): Refere-se à atividade de *criação de substitutos digitais* para objetos do mundo real, ou representá-los digitalmente, seja de forma manual ou automatizada. São exemplos de temas em *Capturing* a digitalização de acervos (RIBEIRO et al., 2020), museus, bibliotecas virtuais e coleções digitais (MEYER; ECCLES, 2016).
- *Creating* (Figura 2.2): Refere-se à atividade de produzir objetos nascidos digitais, que seriam os que são gerados, desenvolvidos e nascidos em meios digitais, seja em forma escrita ou em símbolos, códigos, anotação ou imagem. O estrato *Creating* é ilustrado por temas como escrita, tradução ou outras formas de expressão da linguagem natural ou de máquina, para a criação de códigos, composições, áudios, imagens, design, mídias e desenvolvimento Web (CASTRO, 2020). Num movimento de atravessamento do digital, cuja presença maciça e intensificada nos últimos anos é bem representada pelos identificadores digitais permanentes de autores e publicações conforme exemplifica (FORTES; ALVIM, 2020).
- *Disseminating* (Figura 2.2): Refere-se à atividade de disponibilizar recursos, objetos, resultados de pesquisas ou serviços ao público, podendo incluir publicação, comentários, colaboração, comunicação, direitos autorais e outros.
- *Enriching* (Figura 2.3): Refere-se à adição de informação a um objeto de pesquisa

pela atividade de *anotar, editar ou limpar informações* em projetos de investigação com o objetivo de clarificar a sua origem, natureza, estrutura, significado e seus elementos.

- *Interpreting* (Figura 2.3): Refere-se atividade seguinte a Analyzing cuja função é *explicar e atribuir o significado aos fenômenos observados na análise*, abre o fenômeno ao debate, raciocínio, dedução, indução ou formalização de aspectos específicos.
- *Storing* (Figura 2.4): Refere-se à atividade de fazer cópias digitais sobre o processo de, investigação e resultados, de uma pesquisa e mantê-las acessíveis.

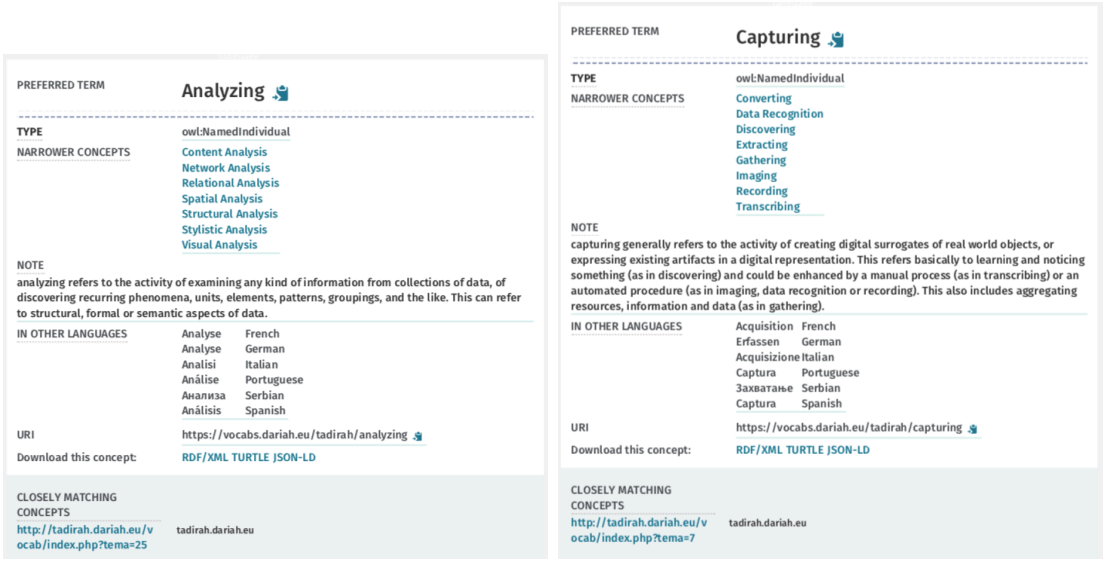


Figura 2.1: Textos descritivos e estratos de segundo nível de *Analyzing* e *Capturing*. Fonte: *TaDiRAH* (2020).

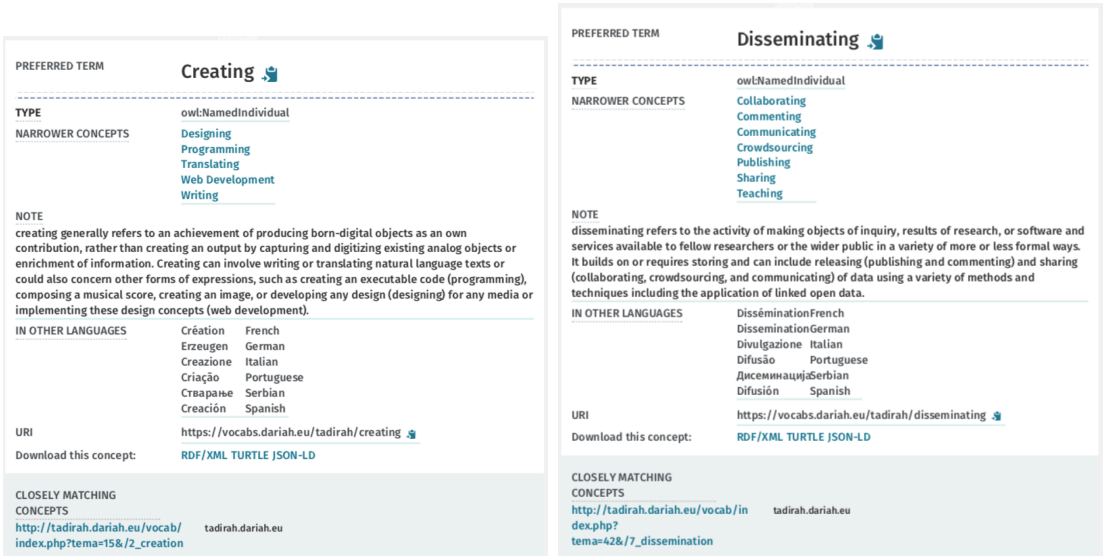


Figura 2.2: Textos descritivos e estratos de segundo nível de *Creating* e *Disseminating*. Fonte: *TaDiRAH* (2020).

Cada estrato desses é subdividido em outros níveis de subestratos sucessivamente menos abrangentes. Dessa forma, os subestratos do segundo nível podem ser vistos como

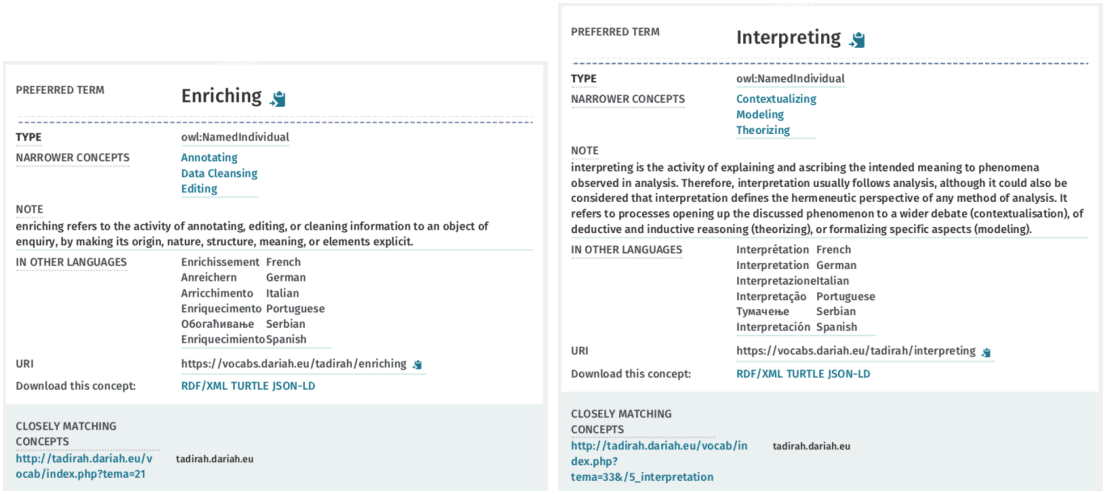


Figura 2.3: Textos descritivos e estratos de segundo nível de *Enriching* e *Interpreting*. Fonte: *TaDiRAH* (2020).

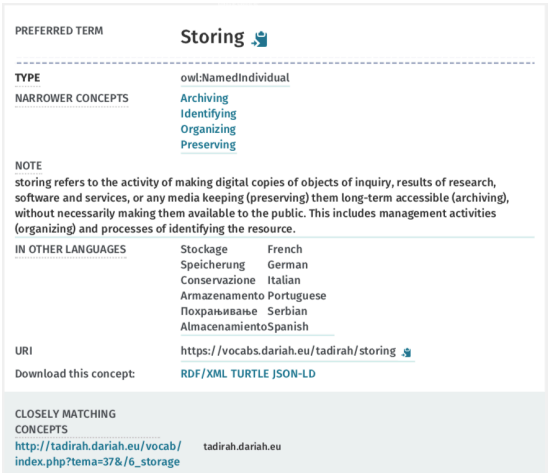


Figura 2.4: Textos descritivos e estratos de segundo nível de *Storing*. Fonte: *TaDiRAH* (2020).

subdivisões dos estratos do primeiro nível. Por sua vez, os subestratos de terceiro nível são subdivisões de subestratos do segundo nível, e assim por diante, caracterizando a natureza hierárquica da taxonomia. Cada estrato ou subestrato é representado por um termo que pode ser usado como palavra-chave associando um documento a um objetivo, uma prática ou um objeto relacionado à pesquisa em Humanidades Digitais. Cada termo ou palavra-chave contém uma descrição textual resumida do que trata cada parte da taxonomia.

Os estratos e subestratos, em seus diferentes níveis, podem ser usados separada ou conjuntamente, segundo a necessidade de cada pesquisador, independentemente de sua localização na estrutura da *TaDiRAH*. A classificação de uma publicação em um estrato deve estar de acordo com o significado dos elementos que identificam o documento, sejam eles palavras-chave ou o seu resumo, como relatam os autores no grupo da (DARIAH-UE CONSORTIUM, 2012).

2.3 Corpus de Textos Descritivos da *TaDiRAH*

Nesta seção, analisamos as relações existentes entre os estratos de primeiro nível da *TaDiRAH* com o objetivo de construir subsídios para as posteriores justificativas dos caminhos adotados no Capítulo 4 e no Capítulo 5. As análises levadas a cabo baseiam-se nas descrições dos estratos apresentados na Seção 2.2.

2.3.1 Sobreposição e Mescla de Estratos

Ao verificar a descrição de *Analyzing*, *Enriching* e *Interpreting* (BOREK et al., 2021), podemos observar dois aspectos relevantes. O primeiro é que as fronteiras entre elas são fluidas, o que é resultado do fato de que muito frequentemente o trabalho do pesquisador de Ciências Humanas envolve uma grande sobreposição de temas. A tentativa de responder a questões como ”O que é análise? O que é agregar informação, em níveis sucessivos na análise? O que é interpretar isso pra gerar um conhecimento novo?”mostra que os temas desses estratos são muito imbricados. O segundo aspecto relevante é que esses são os estratos que envolvem marcadamente a sinergia de métodos analíticos entre Ciências Humanas e Ciência da Computação. A principal consequência é que a classificação de uma quantidade considerável de publicações envolve concomitantemente mais de um dos três estratos.

O grupo de atividades associadas aos estratos *Capturing*, *Creating*, *Disseminating* e *Storing* relaciona-se, a nosso ver, ao trabalho em temas voltados aos aspectos essencialmente tecnológicos das Humanidades Digitais relativos a registro, coleta, armazenamento e recuperação de dados. Nesse sentido, observa-se uma clivagem entre esses estratos e os três estratos abordados anteriormente, *Analyzing*, *Enriching* e *Interpreting*. Os estratos *Capturing* e *Creating* são mesclados devido não somente à correlação temática estreita em atividades de registro e coleta de informações, como também à pequena quantidade de publicações classificadas nesses estratos. Analogamente, os estratos *Disseminating* e *Storing*, cujas atividades incluem eventos de armazenamento (acervo) e a divulgação (acesso) dos trabalhos de pesquisa, são mesclados. Examinamos mais adiante essas relações na seção das sobreposição de estratos.

Essas fronteiras “são fluidas de um lado, de outro lado tem uma certa clivagem”. Existe

Tabela 2.1: Quantidades de textos descritivos dos estratos.

Estrato	Quantidade
Analyzing	25
Capturing-Creating	32
Disseminating-Storing	36
Enriching	17
Interpreting	9

uma divisão clara: "termos que apontam mais para o pesquisador de Ciências Humanas e outros que apontam mais para trabalhos que são de natureza mais técnicas no campo das Humanidades Digitais". Enfim, o imbricamento permite ao mesmo tempo conciliar a classificação de documentos em um campo construído com a convergência de outros campos, com metodologias de construção do conhecimento muito diversificadas, buscando a interdisciplinaridade ao unir múltiplas comunidades específicas interessadas em práticas distintas e no uso de objetos transversais, ao passo que por outro lado demonstra a dificuldade de definir o campo das Humanidades Digitais, onde alguns autores têm se debruçado em duas frentes: (i) uma mais determinista, estabelecendo fronteiras entre o que deve e o que não deve ser considerado Humanidade Digitais e outra mais fluídica em consonância com (DACOS, 2011) onde o potencial de inovação se encontra no entendimento do campo como uma metadisciplina onde a convivência entre as diferentes abordagens, o colaboracionismo em espaços de trocas disciplinares e de experiências analíticas com alta potencialidade de renovação metodológica para as Humanidades e comunicação entre as diversas áreas participantes. É nesta segunda perspectiva que o nosso trabalho se insere.

A opção em classificar documentos de acordo com os estratos de primeiro nível da taxonomia, considerando a mescla de estratos mencionada acima, se dá devido à pequena quantidade de artigos classificados manualmente por especialistas e disponíveis publicamente. Em trabalhos futuros, pode-se adaptar a nossa metodologia para analisar campos específicos das Ciências Humanas, com histórico de vida mais longo, alto volume de produção classificada manualmente por especialistas da área, ou ainda, quando as Humanidades Digitais tiverem muito mais trabalhos manualmente classificados pelos seus especialistas.

2.3.2 Análise dos Textos Descritivos dos Estratos

Na Tabela 2.1 vemos a quantidade de textos descritivos dos estratos extraídos da *TaDiRAH*. Para cada estrato, são consideradas as quantidades de subestratos correspondendo a subdivisões recorrentes desse estrato. A título de ilustração, observamos textos descritivos na Tabela 2.2, na Tabela 2.3 e na Tabela 2.4. Assim sendo, os textos nessas tabelas se referem a uma explicação de cada estrato e de seus substratos. A partir do terceiro nível, há alguns substratos que participam de mais de um estrato de primeiro e segundo níveis como pode ser observado na Tabela 2.5. Quando isso acontece, o estrato mais bem definido, ou seja, que possui a maior quantidade de textos descritivos, podem obter vantagem na hora de serem usados para classificar as publicações.

É possível notar a existência de um destaque de *Analyzing*, *Capturing-Creating* e *Disseminating-Storing* com relação aos demais estratos, destaque esse decorrente da maior quantidade de textos descritivos. Posto de outra forma, os três estratos citados

Tabela 2.2: Textos descritivos originais dos estratos mais abrangentes.

Estrato	**Texto**
Analyzing	analyzing refers to the activity of examining any kind of information from collections of data, of discovering recurring phenomena, units, elements, patterns, groupings, and the like. this can refer to structural, formal or semantic aspects of data.
Capturing-Creating	capturing generally refers to the activity of creating digital surrogates of real world objects, or expressing existing artifacts in a digital representation. this refers basically to learning and noticing something (as in discovering) and could be enhanced by a manual process (as in transcribing) or an automated procedure (as in imaging, data recognition or recording). this also includes aggregating resources, information and data (as in gathering).
Capturing-Creating	creating generally refers to an achievement of producing born-digital objects as an own contribution, rather than creating an output by capturing and digitizing existing analog objects or enrichment of information. creating can involve writing or translating natural language texts or could also concern other forms of expressions, such as creating an executable code (programming), composing a musical score, creating an image, or developing any design (designing) for any media or implementing these design concepts (web development).
Disseminating-Storing	disseminating refers to the activity of making objects of inquiry, results of research, or software and services available to fellow researchers or the wider public in a variety of more or less formal ways. it builds on or requires storing and can include releasing (publishing and commenting) and sharing (collaborating, crowdsourcing, and communicating) of data using a variety of methods and techniques including the application of linked open data.
Enriching	enriching refers to the activity of annotating, editing, or cleaning information to an object of enquiry, by making its origin, nature, structure, meaning, or elements explicit.
Interpreting	interpreting is the activity of explaining and ascribing the intended meaning to phenomena observed in analysis. therefore, interpretation usually follows analysis, although it could also be considered that interpretation defines the hermeneutic perspective of any method of analysis. it refers to processes opening up the discussed phenomenon to a wider debate (contextualisation), of deductive and inductive reasoning (theorizing), or formalizing specific aspects (modeling).
Disseminating-Storing	storing refers to the activity of making digital copies of objects of inquiry, results of research, software and services, or any media keeping (preserving) them long-term accessible (archiving), without necessarily making them available to the public. this includes management activities (organizing) and processes of identifying the resource.

Tabela 2.3: Alguns textos descritivos originais dos substratos de segundo nível.

Estrato	**Substrato**	**Texto**
Enriching	Annotating	annotating refers to the activity of making information about a digital object explicit by adding notes, metadata, keywords, tags or links to a digitized representation or to an annotation file associated with it. this can be in the form of explanatory annotations that comments or contextualize a passage, annotations that make structural or linguistic information explicit, as linked open data making the relationships between objects machine-readable, or in the case of general metadata, adding information about the object as a whole.
Disseminating-Storing	Archiving	archiving includes the process of moving data and other resources to a separate space for retention. if long-term archiving is involved, activities related to data preservation may also be involved.
Disseminating-Storing	Collaborating	collaborating is involved in any research activity being done jointly by several researchers, possibly in different places and at different times. research-oriented collaboration is enabled, particularly, through comprehensive digital research environments, but can also happen around more specific activities, such as communication or sharing of resources.
Disseminating-Storing	Commenting	commenting is the activity of adding information to a piece of data, usually in a way that separates the data to which the comment is attached and the comment. it usually serves to express some opinion, to add contextual information, or to engage in communication or collaboration with others about the object commented on.
Disseminating-Storing	Communicating	communicating refers to the activity of exchanging ideas with other people, primarily, but not exclusively, using linguistic means.
Analyzing	Contentanalysis	content analysis includes researching/studying aspects of objects relating to their meaning, such as identifying concepts or meaningful units.
Interpreting	Contextualizing	contextualizing refers to creating associations between an object of investigation and other, more established or better-understood objects in a relation of geographical, temporal, or thematic proximity to the object of investigation, with the aim of ascribing meaning to that object.

Tabela 2.4: Textos descritivos originais dos substratos de terceiro nível.

Estrato	Substrato	Texto
Capturing-Creating	Gathering	act of collecting together physical or conceptual objects
Capturing-Creating	Discovering	act of searching or traveling around a terrain for the purpose of discovery
Disseminating-Storing	Communicating	act of typing and sending a brief, digital message
Disseminating-Storing	Publishing	action of writing articles or maintaining a weblog
Analyzing	Structuralanalysis	analysis of recurrent co-occurrences of two or more words in language.
Capturing-Creating	Gathering	archiving includes the process of moving data and other resources to a separate space for retention. if long-term archiving is involved, activities related to data preservation may also be involved.
Enriching	Editing	arithmetic operation

beneficiam-se de uma melhor definição dos seus temas e subtemas. Como o auxílio desses textos descritivos com muito mais definições e substratos melhora o entendimento do estrato, a consequência natural é favorecer o uso dessas classificações em contraste com o segundo grupo compartilhado por *Enriching* e *Interpreting*.

Os 20 elementos léxicos mais frequentes nos textos descritivos, elencados na Figura 2.6 com as respectivas frequências de ocorrência, nos mostram a temática transversalizada por todos os estratos: (1) objetos digitais, (2) estudos, pesquisas e métodos, (3) dados e informação. Portanto, não pode-se concluir, a partir das frequências desses elementos léxicos, a predominância de algum estrato sobre os demais (na classificação das publicações) na medida em que os elementos léxicos mais frequentes permeiam diversos estratos. Assim sendo, uma análise mais precisa deve ser capaz de captar, nos temas no Capítulo 4, padrões de associação desses elementos léxicos a temas específicos, que por sua vez também sejam associados a um estrato também específico. Nesse caso, a prevalência de um estrato específico sobre os demais é revelado. Os métodos propostos ao longo dos próximos capítulos abordam essa questão.

2.4 Corpus de Publicações

O corpus de publicações utilizado é constituído de títulos e resumos de artigos e livros em idioma inglês retirados do grupo "*Doing Digital Humanities - A DARIAH Bibliography*" da plataforma *Zotero*. Essas publicações são previamente classificadas artesanalmente por pesquisadores de Humanidades Digitais em cada um dos estratos *Interpreting, Enriching, Analyzing, Capturing, Storing, Disseminating, Creating*, o que adaptamos para os estratos mesclados *Analyzing, Capturing-Creating, Disseminating-Storing, Enriching, Interpreting*. Usamos esses estratos de primeiro nível da taxonomia *TaDiRAH* representados na Tabela 2.7. Naturalmente, em trabalhos futuros, o próprio método poderia ser aplicado sucessivamente aos demais níveis da taxonomia, ou ainda nos substratos de último nível, aqueles que não se desdobram em novos níveis. Desta forma poderemos ter uma classificação sem sobreposições entre as classes.

Tabela 2.5: Algumas publicações em diferentes estratos da *TaDiRAH*.

Texto	**Estrato**	**Substrato**
activity of obtaining information resources relevant to an information need from a collection of information resources	Analyzing, Capturing-Creating	Contentanalysis, Discovering
archiving includes the process of moving data and other resources to a separate space for retention. if long-term archiving is involved, activities related to data preservation may also be involved.	Disseminating-Storing, Capturing-Creating	Archiving, Gathering
binary relation which is left-total (defined on all its input set)	Analyzing, Interpreting, Disseminating-Storing, Enriching	Visualanalysis, Contextualizing, Commenting, Annotating
character string used as a permanent identifier for a digital object, in a format controlled by the international doi foundation	Disseminating-Storing, Analyzing	Identifying, Relationalanalysis,Networkanalysis,Contentanalysis
co-occurrence indicates the semantic proximity or an idiomatic expression of text corpora, based on the statistical-linguistic frequency of occurrence of at least two terms in a certain order.	Interpreting, Disseminating-Storing, Enriching	Contextualizing, Commenting, Annotating
collating describes the activity of comparing two or more versions of written or spoken natural language data by aligning them into alphabetical or numerical order to identify similarities and differences.	Analyzing, Capturing-Creating	Relationalanalysis, Gathering
content on a social media platform, a forum or a blog	Disseminating-Storing, Disseminating-Storing	Sharing, Organizing
contextualizing refers to creating associations between an object of investigation and other, more established or better-understood objects in a relation of geographical, temporal, or thematic proximity to the object of investigation, with the aim of ascribing meaning to that object.	Interpreting, Disseminating-Storing, Enriching	Contextualizing, Commenting, Annotating
data scraping used for extracting data from websites	Capturing-Creating, Capturing-Creating, Disseminating-Storing	Gathering, Gathering, Archiving
extraction of named entity mentions in unstructured text into pre-defined categories	Analyzing, Enriching	Spatialanalysis,Networkanalysis,Contentanalysis, Annotating
identification of something to locations in physical space	Analyzing, Enriching	Spatialanalysis, Annotating
identifying refers to the activity of naming and/or assigning (possibly unique and/or persistent) identifiers such as a word, number, letter, symbol, or any combination to objects of enquiry or to any kind of digital object.	Disseminating-Storing, Analyzing	Identifying, Relationalanalysis,Networkanalysis,Contentanalysis

Tabela 2.6: Maiores frequências de ocorrência nos textos descritivos de estratos.

	Freq
refer	36
object	35
digital	30
activity	28
datum	27
information	24
process	23
text	19
can	18
research	18
analysis	14
language	14
creating	13
system	13
different	12
form	11
method	11
include	11
set	11
data	9

2.4.1 Constituição das Publicações

As informações disponíveis no grupo de publicações já classificadas na plataforma Zotero são as elencadas a seguir. Essas informações são úteis para a determinação das sobreposições de documentos e também como metadados para determinarmos as correlações entre os documentos em um dos métodos de análise que usamos no Capítulo 3.

1. *Item Type*: Tipo de publicação, tais como website, gravação de áudio, postagem em blog, artigo de jornal, revistas ou conferência etc.
2. *Publication Year*: Ano de publicação da obra.
3. *Author*: O principal autor ou criador do trabalho, ou grupo de autores.
4. *doc*: Extensão do arquivo de documento disponível como PDF, Jpeg, MP3 etc.
5. *Publication Title*: Título da publicação.
6. *ISBN*: Sigla no idioma inglês para um identificador internacional único de livros. A Biblioteca Nacional é a organização responsável no Brasil.
7. *ISSN*: Sigla no idioma inglês para identificador único serial de periódicos (jornais e revistas) e trabalhos científicos com oito dígitos. A organização responsável no Brasil é o Centro Brasileiro do ISSN.
8. *DOI*: Sigla em inglês para identificador de qualquer tipo de objetos digitais com um link permanente. Existem diversas organizações que fazem o DOI, como a International DOI Foundation (IDF).
9. *Pages*: O intervalo da página de uma publicação em um periódico maior, por exemplo: *17-27* (da página 17 até a página 27).
10. *Num Pages*: O número de páginas da publicação.
11. *Issue*: Número de um fascículo, suplemento ou edição especial.
12. *Volume*: O número do conjunto onde uma publicação está disponível quando existem diversos blocos ou conjuntos distintos de uma mesma publicação, como em uma enciclopédia.
13. *Series*: Número da série que contém múltiplas publicações como em "Cambridge Studies in Comparative Politics", por exemplo.
14. *Series Number*: O número do item em uma série.
15. *Series Text*: Usado somente para artigos de jornais.
16. *Series Title*: Título de uma série de artigos em um número de um periódico, em uma seção especial ou o título de uma seção.
17. *Publisher*: Pessoa ou empresa que publicou a obra.
18. *Place*: Local de publicação do item.
19. *Language*: Idioma da publicação.
20. *Term*: Termo adicional para identificar itens específicos.
21. *Class*: Estrato ou grupo em que a publicação se encontra.

Tabela 2.7: Publicações classificadas por estrato.

Estrato	Quantidade	Desbalanceamento
Analyzing	236	1:1
Capturing-Creating	24	1:18
Disseminating-Storing	105	1:4
Enriching	43	1:10
Interpreting	35	1:12

```
##        text       datum     digital    analysis       study    research
##         574         380         341         286         270         267
##      method         new         can    literary        work    humanity
##         266         235         229         200         199         196
##        book information      author        word    approach        tool
##         193         176         176         173         164         157
##       paper      result
##         157         143
```

Os 20 elementos léxicos mais frequentes do grupo de publicações no Zotero, mostradas acima, nos dão uma visão holística da temática geral envolvida e o tipo de desafio que enfrentamos durante a classificação. Nessa visão, percebe-se que elementos léxicos mais frequentes nos encaminham para *análise, método, estudo e pesquisa de dados textuais e digitais*. Como consequência, pode-se tomar esse fato como um indicativo de dois aspectos complementares. Primeiro, uma temática geral ligada ao estrato *Analyzing* pode prevalecer sobre os demais e, segundo, o campo de Humanidades Digitais pode ter a tendência de produzir mais trabalhos ligados à temática de *Analyzing* que dos demais estratos. Procedemos à análise de outros indicadores para verificar essa indicação preliminar.

As publicações foram submetidas a um pré-processamento com o intuito limpá-las de quaisquer informações que possam causar ruídos nas análises, preservando as palavras realmente relevantes, as quais são mostradas na Figura 2.5. A limpeza de dados envolveu a retirada de palavras do corpus que, por serem demasiadamente curtas ou frequentes, trariam pouca ou nenhuma contribuição à análise. Essas palavras são normalmente chamadas de *stopwords*. No nosso caso, adicionamos outras às que normalmente são retiradas, são elas: *al, analysis, can, consist, datum, d, de, digital, don, e, e.g, et, etc, f, g, l, le, m, paper, refer, research, use, th, t, text, w, work*. Adicionalmente, um procedimento de determinação de bigramas (elementos léxicos constituídos de duas palavras) é efetuado. Dentre os bigramas determinados, encontram-se *machine_learning, information_retrieval, authorship_attribution, topic_modeling, cluster_analysis, network_analysis, persistent_identifier*.

2.4.2 Distribuição Desbalanceada dos Estratos

Um fato a notar na Tabela 2.7 é a questão do desbalanceamento entre os estratos no que se refere à quantidade de publicações envolvidas no corpus de publicações. Para cada estrato, o desbalanceamento é calculado como a proporção da quantidade de publicações nesse estrato, incluindo mesmo as não exclusivas, frente à soma das quantidades de publicações exclusivas dos demais estratos. As proporções das publicações no corpus de

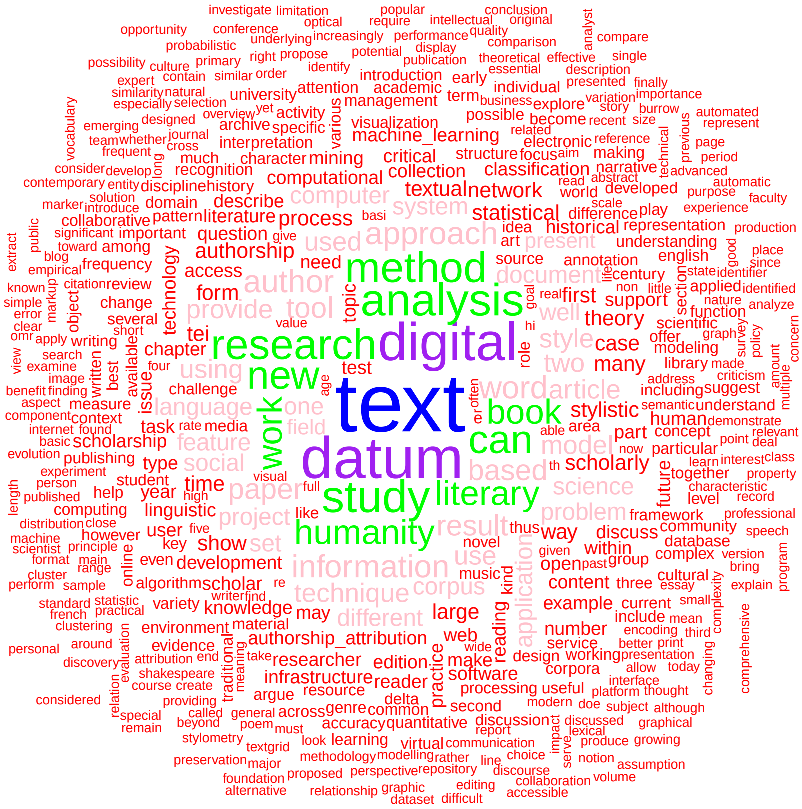


Figura 2.5: Elementos léxicos frequentes nas publicações.

publicações novamente nos levam a pensar que pode ser uma característica do campo a geração de trabalhos mais frequentes no estrato *Analyzing* do que nos demais estratos da *TaDiRAH*. Em uma visão geral, o desbalanceamento entre os estratos é tal que o estrato que tem poucas publicações, tem pouco demais (quase 10 vezes menos), devido ao fato de não termos muitas publicações classificadas, segundo a *TaDiRAH*, por especialistas de Humanidades Digitais (BATISTA; PRATI; MONARD, 2004; DARIAH-UE CONSORTIUM, 2012; KRAWCZYK, 2016).

2.4.3 Sobreposição de Estratos

Uma característica importante do corpus de publicações são os diversos casos explicitamente apontados na Figura 2.6 de múltipla classificação, aquilo que chamamos de sobreposição de estratos. As sobreposições entre as categorias são representados nos gráficos da Figura 2.7, cujo gráfico da esquerda representa, para cada estrato, a quantidade e o percentual das publicações exclusivas e não exclusivas, ou seja, pertencendo a 0, 1 ou 2 outros estratos. Nota-se uma predominância para *Analyzing* e *Disseminating-Storing* em relação à quantidade de textos exclusivos de cada um desses estratos em comparação aos demais. O valor percentual de estratos tem um peso maior nos estratos com poucas publicações como é o caso de *Capturing-Creating* e, principalmente, o de *Interpreting*.

A sobreposição entre pares de estratos é representada no segundo gráfico da Figura 2.7. A porcentagem de cada par de categorias corresponde à quantidade de artigos sobrepostos frente à soma das quantidades individuais desses estratos. Novamente notamos que *Interpreting* possui a maior quantidade de sobreposições, em um nítido contraste com

Figura 2.6: Quantidades de sobreposições entre as publicações dos estratos.

Analyzing e *Disseminating-Storing*.

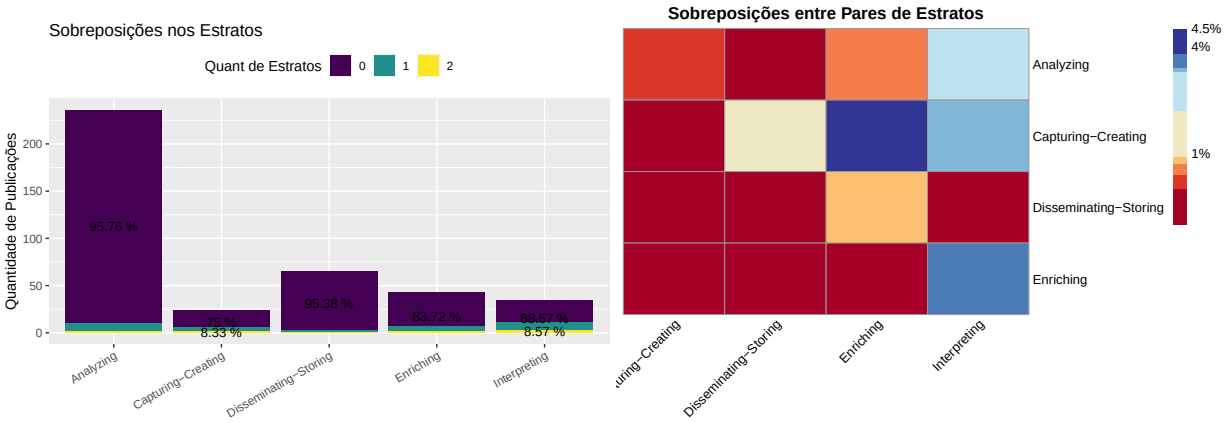


Figura 2.7: Sobreposições entre estratos.

Há três fatores principais que concorrem para essa prevalência temática na produção científica que faz com que os estratos sejam profundamente marcados pela sobreposição. O primeiro deles decorre do processo de informação adotado pela comunidade de pesquisadores e pesquisadoras com fins à teorização e a uma progressiva inteligibilidade dos resultados de suas pesquisas. O segundo fator, por sua vez, decorre da interdisciplinaridade que é uma marca do campo das Humanidades Digitais (DACOS, 2011), o que naturalmente favorece a mescla de temas. Finalmente, o terceiro fator é a forma livre e disjunta de uso da taxonomia como instrumento de classificação, a qual aprofunda a correlação entre os estratos (BOREK et al., 2014, 2016). No entanto, observa-se que características próprias ao processo de classificação do trabalho de pesquisa em Ciências Humanas fazem com que autores omitam alguns dos estratos dentre as palavras-chave (BLOCH, 2002; THOMPSON, 1996).

Também verificamos visualmente a distribuição dos estratos em toda a árvore hierárquica da *TaDiRAH* buscando elementos que possam ser identificados como sobreposição de estratos nas classificações de publicações usando essa taxonomia. Percebemos que existem

sobreposições de estratos na estrutura da taxonomia, não exatamente nos seus estratos de primeiro ou segundo níveis, mas também, com muita frequência, nos de terceiro nível em diante. Para exemplificar escolhemos dois estratos superiores a fim de demonstrar mais detalhadamente: *Analyzing* e *Interpreting* (os dois opostos).

- O estrato *Analyzing* possui 62 substratos dos quais menos da metade (46,77%) são exclusivos, os outros 53,23% são compartilhados com outros estratos a partir do 3º nível dos substratos. O que significa dizer que ao escolhermos aleatoriamente um substrato de *Analyzing* teremos 53,23% de chances de envolver um outro Estrato nesta escolha. Como o estrato *Analyzing* possui mais publicações que os demais, isso não se torna um problema para identificar uma publicação como sendo deste estrato. Porém, o contrário acontece quando se trata de um estrato minoritário, já que as poucas publicações que o estrato minoritário possui ainda tem que compartilhar com um outro estrato que tem muitas publicações associadas, neste caso o maior estrato tende a aparecer mais do que o menor.
- O caso de *Interpreting* é o contrário do de *Analyzing* onde 40% de seus 15 substratos tem ligações com outros estratos, ou seja, há 6 chances dentre 15 sorteios aleatórios de uma publicação de *Interpreting* ser confundida como de outro estrato. Neste caso, se essa publicação for de um estrato majoritário, provavelmente ela será classificada nele e não em *Interpreting*.

Para melhor entendimento, vejamos o substrato *Semantification*, por exemplo, que se refere à realização de processos capazes de identificar e associar informações nos dados para o reconhecimento de ambiguidades e estruturas formais e implícitas para melhorar o reconhecimento, o reuso e a análise dos dados. Se verificarmos em (TADIRAH, 2014), poderemos notar que o substrato *Semantification* provoca sobreposição em três estratos de primeiro nível, isso porque ela compõe substratos de 3º e 4º níveis em estratos diferentes como demonstrado abaixo:

1. *[3º nível de]* Interpreting > Contextualizing > *Semantification*;
2. *[4º nível de]* Enriching > Annotating > Contextualizing > *Semantification*;
3. *[4º nível de]* Disseminating > Commenting > Contextualizing > *Semantification*.

O que significa dizer que se alguma publicação for marcada como *Semantification* pelo seu autor, ela terá sobreposição em 3 estratos: *Interpreting*, *Enriching* e *Disseminating*.

Nos parece lógico afirmar que existe um problema de sobreposição entre os estratos na classificação de publicações segundo a taxonomia *TaDiRAH*, tendo como ponto de partida a própria taxonomia que possui em sua estrutura dualidades entre seus estratos/categorias e substratos/subcategorias. Logo, classificar publicações usando a *TaDiRAH* exige algumas considerações especiais.

3 ANÁLISE AUTOMATIZADA DE TEXTOS

Neste capítulo, descrevemos os métodos de Processamento de Linguagem Natural (denominação abreviada pela sigla *PLN*) que utilizamos neste trabalho, abordando definições e propriedades gerais apontadas na literatura científica como sendo os conceitos essenciais que utilizamos nos métodos propostos nos próximos capítulos. Após um breve preâmbulo no qual apresentamos as definições dos conceitos elementares acerca da análise léxica por tópicos, passamos à descrição de um modelo generativo geral que se assemelha a uma máquina abstrata geradora de textos usando procedimentos probabilísticos (LIU et al., 2016). Em seguida, os três métodos que são objeto deste capítulo são descritos, comparativamente, como especializações do modelo generativo geral. Os métodos específicos apresentados são *Latent Dirichlet Allocation – LDA* (BLEI; NG; JORDAN, 2003), *Correlated Topic Model – CTM* (BLEI; LAFFERTY, 2006b) e *Structural Topic Model - STM* (ROBERTS et al., 2013), dos quais somente o primeiro e o terceiro são elementos constitutivos das propostas desta dissertação. O segundo método é discutido como uma etapa intermediária entre o *LDA* e o *STM*. Terminamos o capítulo abordando ainda o modelo de classificação supervisionada denominado *Supervised Latent Dirichlet Allocation – sLDA* (BLEI; MCAULIFFE, 2010) cujo papel nos capítulos seguintes é o de figurar como base para a comparação dos resultados experimentais.

3.1 Análise Léxica por Tópicos

Em termos bastantes genéricos, um tópico é um tema ou assunto caracterizado por um conjunto de palavras que se relacionam dentro de um ou mais textos. A descoberta de tópicos constitutivos de um conjunto de textos é uma abordagem de PLN baseada em um modelo generativo estatístico geral, modelo este que sorteia de uma forma peculiar palavras segundo seus vínculos estimados com certos tópicos para compor um texto, sem quaisquer considerações semânticas ou sintáticas diretas (GRIFFITHS; STEYVERS, 2004; LUZ, 2013). O ponto de partida é o pressuposto de que o texto é construído como uma combinação em distintas proporções de tópicos no sentido que cada ocorrência de cada palavra está associada a um tópico. Cada tópico, por sua vez, é caracterizado por um vocabulário próprio com cada vocábulo tendo sua própria importância relativa dentro do tópico (BLEI; NG; JORDAN, 2003).

Representar um texto, ou corpus, em forma de uma mescla de tópicos revela as informações nele contidas que não aparecem de imediato por serem muito difíceis de serem encontradas e visualizadas (BLEI, 2012; STEYVERS; GRIFFITHS, 2007). Esses métodos possuem a capacidade de distinguir estruturas ocultas dentro de um corpus (YAU et al., 2014) de forma a expor uma organização temática dos textos que o constituem (VAYANSKY; KUMAR, 2020). Os tópicos também auxiliam na otimização de processos e análises úteis para as pesquisas em grandes quantidades de textos (BLEI, 2012; ROBERTS et al., 2013) na medida em que:

- Facilita generalizações, o gerenciamento de informações e a busca de informações compartilhadas (ALGHAMDI; ALFALQI, 2015; VAYANSKY; KUMAR, 2020);
- Melhora a percepção dos tópicos de maior impacto em um contexto expecífico

(GRIFFITHS; STEYVERS, 2004);

- Auxilia a análise linguística, política e psicológica no âmbito das Ciências Sociais (GRIMMER; STEWART, 2013; STEYVERS; GRIFFITHS, 2007).

3.1.1 Elementos Constitutivos de um Corpus

Passamos aos conceitos e definições essenciais para o entendimento do modelo geral de análise léxica por tópicos. A menor unidade da análise é o denominado *elemento léxico*, definido no Capítulo 2 como uma expressão simples ou composta envolvendo um ou mais vocábulos dotada de um significado específico. A designação de generativo dada ao modelo geral que empregamos indica que esse modelo é uma descrição de um processo de geração de textos. Os propósitos dessa abordagem ficam claros à medida que avançamos na apresentação dos métodos. Por ora, tomemos apenas essa característica generativa do modelo. Assim sendo, a primeira questão que ocorre é quanto a que tipo de objeto é gerado, ou seja, o que é considerado um texto. A resposta a essa questão traz no seu bojo uma primeira simplificação, fato habitual em modelos computacionais: um *texto* é simplesmente uma coleção de elementos léxicos, com as respectivas quantidades de ocorrências. Por sua vez, um *corpus* é definido como uma coleção de textos. Ao conjunto de elementos léxicos que podem ocorrer em um corpus denominamos de *vocabulário*. As tabelas 2.2, 2.3 e 2.4 são ilustrações de textos que, em seu conjunto, formam um corpus, cujo vocabulário é parcialmente mostrado na Tabela 2.6. Em suma, o modelo com o qual trabalhamos é uma máquina abstrata que, empregada sucessivas vezes, constrói, a partir de um vocabulário, os textos constitutivos de um corpus.

A apresentação do modelo generativo geral e dos métodos dele derivados levada adiante no restante do capítulo privilegia o entendimento do funcionamento e a sua interpretação. Com o intuito de tornar essa apresentação precisa, sem no entanto explorar demasiadamente os detalhes matemáticos, alguma notação precisa ser definida com relação a um experimento hipotético de geração de um corpus. O vocabulário empregado no experimento é denotado por $\mathbf{L} = \{\ell_1, \ldots, \ell_L\}$, sendo L a quantidade de elementos léxicos disponíveis nesse vocabulário e no qual todos os elementos léxicos são distintos.

O corpus é dado pelo conjunto $\mathbf{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_D\}$, cuja quantidade de textos é D e no qual cada texto $\mathbf{d}_h = \{w_{1,h}, \ldots, w_{d_h,h}\}$ é uma cesta de d_h elementos léxicos não necessariamente distintos do vocabulário $\mathbf{L}$. A cada elemento léxico $w_{j,h}$ do texto $\mathbf{d}_h$ é associada uma frequência de ocorrência $f_{j,h}$, indicando a quantidade de vezes que o elemento léxico $w_{j,h}$ aparece no documento $\mathbf{d}_h$. A Tabela 3.3 serve de exemplo de definição de um corpus a partir do seu vocabulário e das frequências de ocorrência dos respectivos elementos léxicos nos textos que constituem o corpus. No caso desse exemplo, o corpus usado é constituído dos textos descritivos dos estratos do *TaDiRAH*.

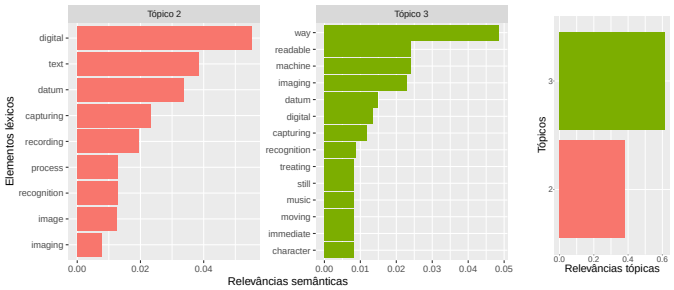
3.1.2 Textos como Mescla de Tópicos em uma Cesta de Elementos Léxicos

A forma como os elementos léxicos são combinados na constituição dos textos que compõem $\mathbf{D}$ confere um perfil semântico próprio a esses textos. No entanto, toma-se como premissa que a ordem de ocorrência não é relevante para determinação indireta de suas propriedades semânticas. À primeira vista, a hipótese de que o perfil semântico de um texto seja definido pelas diversas ocorrências dos elementos léxicos que o compõem, sendo

irrelevante a ordem em que esses elementos léxicos ocorrem pode parecer demasiadamente forte. No entanto, essa hipótese tem se mostrado adequada para fins de análise hermenêutica em inúmeros casos reais, sobretudo quando o escopo dos textos analisados é relativamente restrito (BLEI; NG; JORDAN, 2003). Por outro lado, considerar cada texto como uma cesta de elementos léxicos é apropriado para a utilização de ferramentas estatísticas que traçam um perfil das ocorrências simultâneas dos elementos léxicos que, indiretamente, revelam propriedades semânticas. Essa questão será retomada mais adiante.

Antes da discussão sobre os detalhes do funcionamento do modelo generativo que servem de base para as ferramentas estatísticas, enunciamos uma segunda premissa, segundo a qual cada texto é constituído por uma mescla de uma quantidade fixa T de *tópicos*, os quais formam o conjunto $\mathbf{T} = \{t_1, \ldots, t_T\}$. Cada tópico t_j em $\mathbf{T}$ é uma função que expressa um movimento temático de $\mathbf{L}$ por meio do estabelecimento da relevância de cada elemento léxico dentro do tópico (ALGHAMDI; ALFALQI, 2015). Sendo assim, cada tópico t_j consiste em uma função massa de probabilidade $\boldsymbol{\beta}_j$, atribuindo uma probabilidade de ocorrência $\beta_{j,i}$, denominada *relevância semântica*, a cada elemento léxico ℓ_i do vocabulário (VAYANSKY; KUMAR, 2020). Cabe observar que as relevâncias semânticas dependem unicamente dos membros de $\mathbf{L}$ e são independentes dos textos. Se o agrupamento de elementos léxicos através das respectivas relevâncias semânticas confere uma interpretação objetiva a um tópico, a adequada combinação de tópicos é uma forma de descrever um texto. Nesse sentido, usamos a *relevância tópica* $\theta_{j,h}$ como sendo a proporção em que t_j ocorre em um texto $\mathbf{d}_h$.

Na ilustração a seguir, vê-se um exemplo de tópicos envolvidos em um texto descritivo de estrato da *TaDiRAH* (segundo a análise com o método LDA, detalhado na sequência do capítulo). Cada tópico é composto por vários elementos léxicos, os quais são distribuídos dentro de uma publicação em uma proporção definida por uma combinação dos tópicos a que as ocorrências estão associadas. Nesse exemplo, o vocabulário de cada tópico é definido pelos elementos léxicos de $\mathbf{L}$ que contribuem para a interpretação objetiva desse tópico, ou seja, os elementos léxicos cujas relevâncias semânticas no tópico são maiores que zero. Alguns elementos léxicos que aparecem originalmente no plural no texto descritivo pode ter a sua versão no singular associada a um tópico.



Texto Descritivo:
Capturing-Creating_Datarecognition.1

data recognition, e.g. optical character recognition (ocr), refers to the process of treating the immediate products of digital data capturing (recording or imaging), such as digital facsimiles of texts or of music score, in a way to extract discrete, machine readable units from them, such as plain text words, musical notes, still or moving image elements.

A proporção de ocorrência de cada elemento léxico em uma publicação depende das relevâncias tópicas dos tópicos que formam essa publicação, assim como da sua relevância semântica (a forma empregada) nos tópicos. Ou ainda, para cada elemento léxico no tópico, a sua frequência será corresponde a sua relevância no tópico e a forma como está sendo usado para expressar ideias. Sendo assim, também podemos afirmar que quanto maior a frequência do elemento léxico no tópico, menor será o uso de sinônimos que expressam a mesma ideia neste mesmo tópico, e maior será a sua relevância semântica no tópico. Cabe também ressaltar que diferentes ocorrências de um elemento léxico presente em uma mesma publicação podem estar associadas a diferentes tópicos.

3.1.3 Modelo para Análise por Tópicos

O modelo generativo pode ser visto como uma máquina abstrata que joga dados que gera cada texto individualmente através do sorteio de elementos léxicos sem ordem de colocação entre eles, formando assim a cesta de elementos léxicos correspondente. A análise por tópicos parte da premissa adicional de que o corpus sob análise é uma amostra do que essa máquina abstrata pode gerar e da qual são extraídos os padrões semânticos (VAYANSKY; KUMAR, 2020). Assim sendo, os padrões semânticos do corpus podem ser extraídos diretamente da forma de funcionamento do modelo generativo cujos princípios estão descritos nas subseções anteriores. Alguns dos elementos de análise são diretamente observáveis a partir do corpus, como a quantidade d_h de elementos léxicos no texto $\mathbf{d}_h$, e outros são latentes, notadamente as relevâncias semântica e tópica. Para descrever o processo na construção de $\mathbf{d}_h$ nesse modelo, observamos que cada um dos d_h elementos léxicos é gerado aleatoriamente, segundo a sua relevância semântica, associado a um tópico, o qual também é escolhido aleatoriamente segundo a relevância tópica.

As relevâncias tópicas são derivadas de uma estratégia de atribuição de tópicos que especifica o método empregado, sendo assunto das próximas subseções. Analogamente ao comentário em (BLEI; NG; JORDAN, 2003), supomos a distribuição de Poisson para a determinação do valor de d_h de cada texto $\mathbf{d}_h$, embora não seja essencial visto que outras distribuições podem também ser usadas. O essencial a ser notado é que tomamos

d_h como sendo independente de todas as outras distribuições envolvidas. Uma descrição resumida do modelo generativo para cada texto $\mathbf{d}_h$, tendo sido a quantidade de tópicos T pré-estabelecida e as relevâncias semânticas $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T$ escolhidas segundo uma estratégia de atribuição de elementos léxicos, consiste em uma sequência de sorteios de tópico z segundo $\boldsymbol{\theta}_h$ e de elementos léxicos segundo β_z repetidos d_h vezes, ou seja:

Algoritmo: MODELOGENERATIVO(T)

1: Criar a cesta de elementos léxicos vazia
2: Escolher quantidade de elementos léxicos d_h
3: Determinam-se as relevâncias tópicas $\boldsymbol{\theta}_h$ para T tópicos
4: Para n valendo $1, \ldots, d_h$, realizar os seguintes passos:
5: Escolher tópico z_n segundo $\boldsymbol{\theta}_h$ (formalmente, $z_n \sim Multinomial(\boldsymbol{\theta}_h)$)
6: Escolher elemento léxico ℓ_n usando a relevância semântica β_{z_n} condicionada pelo tópico z_n
7: Acrescentar ℓ_n à cesta de elementos léxicos

Como ilustração do princípio utilizado no modelo generativo, tomamos uma vez mais uma parte do corpus dos textos descritivos dos estratos da *TaDiRAH* descritos no Capítulo 2 (ALGHAMDI; ALFALQI, 2015). Os textos escolhidos para esse exemplo são aqueles que constam da Tabela 3.1, para os quais são exibidas as relevâncias tópicas de um modelo com 7 tópicos. Seguindo a hipótese de que o corpus constituído pelas cestas de elementos léxicos descritos na Tabela 3.3 é uma amostra do que o modelo generativo descrito no Algoritmo MODELOGENERATIVO(T) pode gerar com $T = 7$ tópicos, observamos, por exemplo, que a geração do texto *Capturing-Creating_Datarecognition.1* emprega a relevância semântica $\boldsymbol{\beta}_3$ em 61,48% de seus elementos léxicos. Esse fato é uma consequência da escolha na linha 3 ser realizada com as probabilidades indicadas na linha correspondente a $\boldsymbol{\theta}_{Capturing-Creating\_Datarecognition.1}$ na Tabela 3.1. Por sua vez, as relevâncias semânticas na Tabela 3.2 indicam que o elemento léxico *analysis* é escolhido na linha 6 em 4,85% das vezes em que $\boldsymbol{\beta}_3$ é empregada.

Tabela 3.1: Relevância tópica em um corpus formado por textos descritivos de estratos da *TaDiRAH*.

	Tópico 1	**Tópico 2**	**Tópico 3**	**Tópico 4**	**Tópico 5**	**Tópico 6**	**Tópico 7**
$\theta_{Analyzing\_Analyzing.1}$	0.0011	0.0011	0.0011	0.0011	0.9934	0.0011	0.0011
$\theta_{Analyzing\_Contentanalysis.1}$	0.0015	0.0015	0.9912	0.0015	0.0015	0.0015	0.0015
$\theta_{Analyzing\_Contentanalysis.2}$	0.0016	0.0016	0.0016	0.8687	0.0016	0.1234	0.0016
$\theta_{Analyzing\_Networkanalysis.1}$	0.0013	0.0013	0.0013	0.0013	0.3011	0.0013	0.6927
$\theta_{Analyzing\_Relationalanalysis.1}$	0.0008	0.0008	0.0008	0.0008	0.9952	0.0008	0.0008
$\theta_{Analyzing\_Spatialanalysis.1}$	0.0013	0.0013	0.5186	0.0013	0.4752	0.0013	0.0013
$\theta_{Analyzing\_Structuralanalysis.1}$	0.0017	0.0017	0.0017	0.0017	0.9895	0.0017	0.0017
$\theta_{Analyzing\_Stylisticanalysis.1}$	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.9925
$\theta_{Analyzing\_Stylisticanalysis.2}$	0.9934	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011
$\theta_{Analyzing\_Stylisticanalysis.3}$	0.0011	0.0011	0.0011	0.0011	0.9934	0.0011	0.0011
$\theta_{Analyzing\_Stylisticanalysis\_Relationalanalysis.1}$	0.0016	0.0016	0.0016	0.0016	0.9904	0.0016	0.0016
$\theta_{Analyzing\_Stylisticanalysis\_Relationalanalysis.2}$	0.0013	0.0013	0.0013	0.0013	0.0013	0.8702	0.1231
$\theta_{Analyzing\_Stylisticanalysis\_Relationalanalysis.3}$	0.0016	0.0016	0.0016	0.0016	0.0016	0.9904	0.0016
$\theta_{Analyzing\_Visualanalysis.1}$	0.0007	0.0007	0.0007	0.0007	0.9956	0.0007	0.0007
$\theta_{Capturing-Creating\_Capturing.1}$	0.0006	0.9962	0.0006	0.0006	0.0006	0.0006	0.0006
$\theta_{Capturing-Creating\_Converting.1}$	0.0005	0.9969	0.0005	0.0005	0.0005	0.0005	0.0005
$\theta_{Capturing-Creating\_Converting.2}$	0.0015	0.0015	0.0015	0.9912	0.0015	0.0015	0.0015
$\theta_{Capturing-Creating\_Creating.1}$	0.0005	0.9972	0.0005	0.0005	0.0005	0.0005	0.0005
$\theta_{Capturing-Creating\_Datarecognition.1}$	0.0007	0.3816	0.6148	0.0007	0.0007	0.0007	0.0007
$\theta_{Capturing-Creating\_Designing.1}$	0.0006	0.0006	0.0006	0.9964	0.0006	0.0006	0.0006
$\theta_{Capturing-Creating\_Discovering.1}$	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.9950
$\theta_{Capturing-Creating\_Extracting.1}$	0.0006	0.0006	0.0006	0.0006	0.0006	0.9965	0.0006
$\theta_{Capturing-Creating\_Gathering.1}$	0.0012	0.0012	0.0012	0.0012	0.0012	0.9930	0.0012
$\theta_{Capturing-Creating\_Imaging.1}$	0.0009	0.0009	0.9947	0.0009	0.0009	0.0009	0.0009
$\theta_{Capturing-Creating\_Programming.1}$	0.9944	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009
$\theta_{Capturing-Creating\_Recording.1}$	0.9938	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
$\theta_{Capturing-Creating\_Transcribing.1}$	0.0009	0.9947	0.0009	0.0009	0.0009	0.0009	0.0009
$\theta_{Capturing-Creating\_Translating.1}$	0.0016	0.0016	0.0016	0.9904	0.0016	0.0016	0.0016
$\theta_{Capturing-Creating\_Writing.1}$	0.0013	0.8714	0.0013	0.0013	0.0013	0.0013	0.1219
$\theta_{Disseminating-Storing\_Archiving\_Gathering.1}$	0.9934	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011
$\theta_{Disseminating-Storing\_Collaborating.1}$	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.9954
$\theta_{Disseminating-Storing\_Commenting.1}$	0.0009	0.0009	0.0009	0.0009	0.0009	0.9947	0.0009
$\theta_{Disseminating-Storing\_Commenting.2}$	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.9934
$\theta_{Disseminating-Storing\_Crowdsourcing.1}$	0.0010	0.0010	0.0010	0.0010	0.9938	0.0010	0.0010
$\theta_{Disseminating-Storing\_Disseminating.1}$	0.9962	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
$\theta_{Disseminating-Storing\_Organizing\_Collaborating.1}$	0.3468	0.0013	0.6470	0.0013	0.0013	0.0013	0.0013
$\theta_{Disseminating-Storing\_Preserving\_Gathering.1}$	0.7994	0.0012	0.1948	0.0012	0.0012	0.0012	0.0012
$\theta_{Disseminating-Storing\_Publishing.1}$	0.9961	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
$\theta_{Disseminating-Storing\_Storing.1}$	0.9952	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008
$\theta_{Enriching\_Annotating.1}$	0.0004	0.0004	0.0004	0.0004	0.0004	0.9975	0.0004
$\theta_{Enriching\_Annotating.2}$	0.0007	0.0007	0.0007	0.9956	0.0007	0.0007	0.0007
$\theta_{Enriching\_Datacleansing.1}$	0.0015	0.9912	0.0015	0.0015	0.0015	0.0015	0.0015
$\theta_{Enriching\_Editing.1}$	0.4374	0.0016	0.0016	0.0016	0.0016	0.5547	0.0016
$\theta_{Interpreting\_Interpreting.1}$	0.0008	0.0008	0.0008	0.9950	0.0008	0.0008	0.0008
$\theta_{Interpreting\_Modeling.1}$	0.0005	0.0005	0.9969	0.0005	0.0005	0.0005	0.0005
$\theta_{Interpreting\_Theorizing.1}$	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.9947
$\theta_{Interpreting\_Theorizing.2}$	0.0007	0.0007	0.9956	0.0007	0.0007	0.0007	0.0007

Tabela 3.2: Relevância semântica de alguns elementos léxicos em uma análise com 7 tópicos.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
object	0.0277	0.0328	0.0299	0.0188	0.0158	0.0484	0.0187
activity	0.0453	0.0185	0.0117	0.0094	0.0158	0.0217	0.0391
digital	0.0233	0.0551	0.0134	0.0000	0.0158	0.0221	0.0094
can	0.0300	0.0131	0.0162	0.0188	0.0260	0.0072	0.0253
research	0.0540	0.0000	0.0000	0.0000	0.0079	0.0000	0.0655
information	0.0000	0.0262	0.0000	0.0000	0.0158	0.0723	0.0094
datum	0.0204	0.0336	0.0150	0.0094	0.0200	0.0223	0.0000
text	0.0000	0.0383	0.0001	0.0086	0.0394	0.0082	0.0197
analysis	0.0000	0.0000	0.0081	0.0373	0.0473	0.0074	0.0000
language	0.0000	0.0066	0.0000	0.0466	0.0158	0.0147	0.0000
method	0.0120	0.0000	0.0081	0.0188	0.0180	0.0053	0.0186
way	0.0000	0.0000	0.0485	0.0000	0.0000	0.0219	0.0000
different	0.0120	0.0131	0.0081	0.0094	0.0079	0.0000	0.0187
include	0.0251	0.0197	0.0228	0.0000	0.0000	0.0000	0.0000
creating	0.0060	0.0525	0.0081	0.0000	0.0000	0.0000	0.0000
process	0.0180	0.0127	0.0000	0.0281	0.0000	0.0077	0.0000
data	0.0172	0.0000	0.0117	0.0000	0.0291	0.0072	0.0000
form	0.0000	0.0066	0.0244	0.0000	0.0000	0.0217	0.0093
may	0.0102	0.0131	0.0349	0.0000	0.0000	0.0000	0.0000
making	0.0272	0.0000	0.0000	0.0000	0.0079	0.0178	0.0000
aspect	0.0000	0.0066	0.0126	0.0188	0.0114	0.0000	0.0000
usually	0.0000	0.0000	0.0162	0.0094	0.0000	0.0217	0.0000
structural	0.0000	0.0000	0.0081	0.0000	0.0236	0.0145	0.0000
user	0.0000	0.0000	0.0000	0.0375	0.0079	0.0000	0.0000
linguistic	0.0000	0.0000	0.0000	0.0279	0.0079	0.0074	0.0000
involve	0.0060	0.0197	0.0000	0.0094	0.0079	0.0000	0.0000
natural	0.0000	0.0066	0.0000	0.0279	0.0000	0.0074	0.0000
mean	0.0000	0.0000	0.0081	0.0000	0.0079	0.0072	0.0187
representation	0.0000	0.0262	0.0081	0.0000	0.0000	0.0072	0.0000
capturing	0.0060	0.0234	0.0117	0.0000	0.0000	0.0000	0.0000
source	0.0120	0.0000	0.0000	0.0281	0.0000	0.0000	0.0000
machine	0.0000	0.0000	0.0240	0.0000	0.0000	0.0148	0.0000
readable	0.0000	0.0000	0.0240	0.0000	0.0000	0.0148	0.0000
specific	0.0120	0.0000	0.0000	0.0094	0.0079	0.0000	0.0094
set	0.0000	0.0000	0.0000	0.0000	0.0315	0.0072	0.0000
network	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0374
existing	0.0000	0.0260	0.0000	0.0000	0.0000	0.0000	0.0097
phenomenon	0.0000	0.0000	0.0162	0.0188	0.0000	0.0000	0.0000
modeling	0.0000	0.0000	0.0244	0.0094	0.0000	0.0000	0.0000
images	0.0000	0.0000	0.0162	0.0000	0.0079	0.0000	0.0094
element	0.0000	0.0000	0.0162	0.0000	0.0079	0.0000	0.0094
open	0.0180	0.0000	0.0000	0.0000	0.0079	0.0072	0.0000
model	0.0000	0.0000	0.0325	0.0000	0.0000	0.0000	0.0000
formal	0.0060	0.0000	0.0000	0.0094	0.0158	0.0000	0.0000
imaging	0.0000	0.0077	0.0229	0.0000	0.0000	0.0000	0.0000
result	0.0240	0.0066	0.0000	0.0000	0.0000	0.0000	0.0000
analyzing	0.0060	0.0000	0.0000	0.0094	0.0079	0.0072	0.0000
writing	0.0000	0.0125	0.0000	0.0000	0.0000	0.0060	0.0118
relation	0.0000	0.0066	0.0000	0.0000	0.0236	0.0000	0.0000
describe	0.0060	0.0066	0.0081	0.0000	0.0000	0.0000	0.0094

Tabela 3.3: Quantidades de ocorrências nas cestas de elementos léxicos correspondentes a textos descritivos dos estratos da *TaDiRAH*.

	Analyzing_Analyzing.1	Enriching_Annotating.1	Analyzing_Stylisticanalysis.1	Capturing-Creating_Capturing.1	Disseminating-Storing_Collaborating.1	Disseminating-Storing_Commenting.1	Interpreting_Theorizing.1	Analyzing_Contentanalysis.1	Analyzing_Stylisticanalysis_Relationalanalysis.1	Capturing-Creating_Converting.1	Capturing-Creating_Creating.1	Disseminating-Storing_Crowdsourcing.1	Enriching_Datacleansing.1
act	0	0	0	0	0	0	0	0	0	0	0	0	0
collecting	0	0	0	0	0	0	0	0	0	0	0	0	0
conceptual	0	0	0	0	0	0	0	0	0	0	0	0	0
object	0	3	0	0	0	1	1	1	0	2	2	0	1
traveling	0	0	0	0	0	0	0	0	0	0	0	0	0
terrain	0	0	0	0	0	0	0	0	0	0	0	0	0
purpose	0	0	0	0	0	0	0	0	0	0	0	0	0
discovery	0	0	0	0	0	0	0	0	0	0	0	0	0
typing	0	0	0	0	0	0	0	0	0	0	0	0	0
brief	0	0	0	0	0	0	0	0	0	0	0	0	0
digital	0	1	0	2	1	0	0	0	0	3	1	1	1
writing	0	0	0	0	0	0	0	0	0	0	1	0	0
article	0	0	0	0	0	0	0	0	0	0	0	0	0
maintaining	0	0	0	0	0	0	0	0	0	0	0	0	0
weblog	0	0	0	0	0	0	0	0	0	0	0	0	0
analysis	0	0	0	0	0	0	0	1	0	0	0	0	0
recurrent	0	0	0	0	0	0	0	0	0	0	0	0	0
co	0	0	0	0	0	0	0	0	0	0	0	0	0
two	0	0	0	0	0	0	0	0	0	0	0	0	0
word	0	0	0	0	0	0	0	0	0	0	0	0	0
language	0	0	0	0	0	0	0	0	0	0	1	0	0
analyzing	1	0	0	0	0	0	0	0	0	0	0	0	0
activity	1	1	0	1	1	1	1	0	0	0	0	0	0
examining	1	0	0	0	0	0	0	0	0	0	0	0	0
kind	1	0	0	0	0	0	0	0	0	0	0	0	0
information	1	3	0	1	0	2	0	0	0	0	1	0	0
collection	1	0	0	0	0	0	0	0	0	0	0	0	0
data	2	0	0	0	0	1	0	0	0	0	0	0	0
recurring	1	0	0	0	0	0	0	0	0	0	0	0	0
phenomena	1	0	0	0	0	0	0	0	0	0	0	0	0
groupings	1	0	0	0	0	0	0	0	0	0	0	0	0
like	1	0	0	0	0	0	0	0	0	0	0	0	0
can	1	1	0	0	1	0	0	0	0	0	1	0	0
structural	1	1	0	0	0	0	0	0	0	0	0	0	0
formal	1	0	0	0	0	0	0	0	0	0	0	0	0
aspect	1	0	0	0	0	0	0	1	0	0	0	0	0
annotating	0	1	0	0	0	0	0	0	0	0	0	0	0
making	0	2	0	0	0	0	0	0	0	0	0	0	0
explicit	0	2	0	0	0	0	0	0	0	0	0	0	0
adding	0	2	0	0	0	1	0	0	0	0	0	0	0
metadata	0	2	0	0	0	0	0	0	0	1	0	0	0
keywords	0	1	0	0	0	0	0	0	0	0	0	0	0
tag	0	1	0	0	0	0	0	0	0	0	0	0	0
link	0	1	0	0	0	0	0	0	0	0	0	0	0
digitized	0	1	0	0	0	0	0	0	0	0	0	0	0
representation	0	1	0	1	0	0	0	0	0	1	0	0	0
annotation	0	3	0	0	0	0	0	0	0	0	0	0	0
file	0	1	0	0	0	0	0	0	0	2	0	0	0
associated	0	1	0	0	0	0	0	0	0	0	0	0	0
form	0	1	0	0	0	0	0	0	0	0	1	0	0
explanatory	0	1	0	0	0	0	0	0	0	0	0	0	0
comment	0	1	0	0	0	2	0	0	0	0	0	0	0
contextualize	0	1	0	0	0	0	0	0	0	0	0	0	0
passage	0	1	0	0	0	0	0	0	0	0	0	0	0
linguistic	0	1	0	0	0	0	0	0	0	0	0	0	0
linked	0	1	0	0	0	0	0	0	0	0	0	0	0
open	0	1	0	0	0	0	0	0	0	0	0	1	0
datum	0	1	0	2	0	1	0	0	0	1	0	0	1
machine	0	1	0	0	0	0	0	0	0	0	0	0	0
readable	0	1	0	0	0	0	0	0	0	0	0	0	0
general	0	1	0	0	0	0	0	0	0	0	0	0	0
whole	0	1	0	0	0	0	0	0	0	0	0	0	0
arithmetic	0	0	0	0	0	0	0	0	0	0	0	0	0
operation	0	0	0	0	0	0	0	0	0	0	0	0	0
art	0	0	0	0	0	0	0	0	0	0	0	0	0
science	0	0	0	0	0	0	0	0	0	0	0	0	0
creating	0	0	0	1	0	0	0	0	0	0	5	0	0
image	0	0	0	0	0	0	0	0	0	0	1	0	0
recording	0	0	0	1	0	0	0	0	0	0	0	0	0
authorship	0	1	0	0	0	0	0	0	0	0	0	0	0
attribution	0	1	0	0	0	0	0	0	0	0	0	0	0
mean	0	0	1	0	0	0	1	0	0	0	0	0	0
theory	0	0	1	0	0	0	0	0	0	0	0	0	0
method	0	0	1	0	0	0	0	0	0	0	0	0	0
determining	0	0	1	0	0	0	0	0	0	0	0	0	0
author	0	0	1	0	0	0	0	0	0	0	0	0	0
piece	0	0	1	0	0	1	0	0	0	0	0	0	0
work	0	0	1	0	0	0	0	0	0	0	0	0	0
text	0	0	1	0	0	0	1	0	0	0	1	0	1
appropriate	0	0	1	0	0	0	0	0	0	0	0	0	0
characterization	0	0	1	0	0	0	0	0	0	0	0	0	0
stylistic	0	0	1	0	0	0	0	0	0	0	0	0	0
thematic	0	0	1	0	0	0	0	0	0	0	0	0	0
branch	0	0	0	0	0	0	0	0	0	0	0	0	0
machine_learning	0	0	0	0	0	0	0	0	0	0	0	0	0
statistic	0	0	0	0	0	0	0	0	0	0	0	0	0
computer	0	0	0	0	0	0	0	0	0	0	0	0	0
study	0	0	0	0	0	0	0	0	0	0	0	0	0
algorithm	0	0	0	0	0	0	0	0	0	0	0	0	0
architecture	0	0	0	0	0	0	0	0	0	0	0	0	0
learn	0	0	0	0	0	0	0	0	0	0	0	0	0
fact	0	0	0	0	0	0	0	0	0	0	0	0	0
broadcast	0	0	0	0	0	0	0	0	0	0	0	0	0
medium	0	0	0	0	0	0	0	0	0	0	0	0	0
capturing	0	0	0	1	0	0	0	0	0	0	1	0	0
generally	0	0	0	1	0	0	0	0	0	0	1	0	0
surrogate	0	0	0	1	0	0	0	0	0	0	0	0	0
real	0	0	0	1	0	0	0	0	0	0	0	0	0
objects	0	0	0	1	0	0	0	0	0	0	0	0	0
expressing	0	0	0	1	0	0	0	0	0	0	0	0	0

Esse exemplo que inclui as relevâncias tópicas na Tabela 3.1 e as relevâncias semânticas na Tabela 3.2 ilustra o funcionamento da modelagem descrita no Algoritmo MODELOGENERATIVO($T = 7$) como uma máquina abstrata que imita o perfil de escrita usado por um determinado conjunto de autores ao criar os textos originais vistos como as cestas de elementos léxicos da Tabela 3.3.

Os métodos específicos *LDA* (BLEI; NG; JORDAN, 2003), CTM (BLEI; LAFFERTY, 2006b) e STM (ROBERTS et al., 2013), são especializações do modelo generativo que se diferenciam pela forma como são construídos as relevâncias tópicas, temas das subseções seguintes. O objetivo de cada um desses métodos na análise por tópicos de um corpus é reconstruir, a partir dos textos do corpus, os parâmetros do modelo generativo que melhor descreve os textos sendo analisados, notadamente as suas relevâncias tópicas e semânticas. Para atingir esse objetivo, são empregados métodos estatísticos de análise e cálculo computacional, métodos esses que estão fora do escopo desta dissertação. Nesse sentido, as próximas subseções, relativas aos três métodos específicos de análise por tópicos, tratam apenas da exposição dos aspectos específicos do modelo generativo geral.

3.2 Atribuição de Tópicos por *Dirichlet*

VAYANSKY; KUMAR (2020) consideram que o *LDA* (BLEI; NG; JORDAN, 2003) é o desfecho de um movimento de modelagem léxica por tópicos que começou com três de seus antecessores, a saber (DEERWESTER et al., 1990; HOFMANN, 2017; SALTON, 1983), tornando-se referência para as diversas variações que o seguiram. Essa tendência prevalece ainda hoje embora haja modelos propostos recentemente que adotam outras formas de codificação do vocabulário capazes de capturar certas propriedades sintáticas ou semânticas, como os vetores em um espaço multidimensional em (COSTA; ORTALE, 2021). As características específicas ao método *LDA* que o diferem daqueles que serão descritos na sequência dizem respeito essencialmente à estratégia de determinação das relevâncias tópicas. Os principais elementos de especificação do *LDA* quanto ao modelo generativo geral são as seguintes:

1. *Relevâncias semânticas fixas para todos os textos*: todos são gerados com as mesmas relevâncias semânticas, ou seja, a relevância semântica β_{z_n} usada na linha 6 independe do texto sendo gerado. Esta é uma propriedade comum aos três métodos descritos neste capítulo.

2. *Relevâncias tópicas por Dirichlet*: conforme indicado na linha 3, as relevâncias tópicas $\boldsymbol{\theta}_h$ que governam a atribuição de tópicos na geração do texto $\mathbf{d}_h$ lhe são próprias. A particularidade do *LDA* está na estratégia usada para determinar essas relevâncias tópicas, as quais são escolhidas a partir de um parâmetro específico em uma distribuição de Dirichlet. O parâmetro empregado para especificar a distribuição, o qual determina o perfil dos textos gerados, é único para todos os textos do corpus.

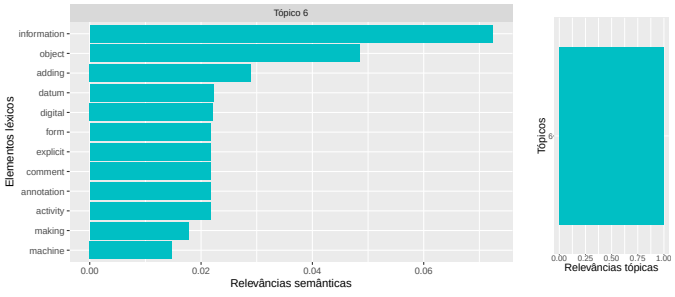
3. *Não há correlação entre tópicos*: uma característica do *LDA* que decorre da adoção da estratégia baseada na distribuição de Dirichlet é o fato de o modelo não considerar as possíveis correlações entre as relevâncias tópicas nos tópicos gerados. Esta é uma característica que o difere dos demais métodos deste capítulo.

A distribuição Dirichlet pode ser entendida como uma distribuição de distribuições, ou seja, ela gera distribuições. A existência desse parâmetro confere flexibilidade ao modelo,

pois o seu comportamento pode assim ser calibrado, podendo gerar dados mais inclinados a poucos tópicos ou mais igualitários sobre os tópicos. Esse parâmetro dá ao modelo uma infinidade de formas diferentes de viciar os dados de tópicos pois correspondem a todos os vetores de tamanho igual à quantidade de tópicos de valores que somam 1. O ajuste no parâmetro da distribuição de Dirichlet depende do valor de T de tal forma que quanto maior esse valor, maior será a participação de tópicos nas publicações (BLEI; NG; JORDAN, 2003; VAYANSKY; KUMAR, 2020). Vejamos dois cenários possíveis.

Cenário dado mais igualitário quer dizer que a distribuição favorece mais as chances mais igualitárias entre os tópicos, o que privilegia o sorteio de uma combinação mais diversificada de vários tópicos. Então, quando você gera uma publicação com um dado desse tipo, você tende a misturar muitos tópicos, porque todos eles têm chances parecidas de ocorrer. Múltiplos tópicos são influentes em *Capturing-Creating_Datarecognition.1*, sobretudo 3 e 2. Nesse caso, 38,16% dos elementos léxicos são gerados com $\boldsymbol{\beta}_2$.

Cenário dado menos igualitário com o dado mais inclinado a um tópico, indica que esse tópico prevalecerá sobre os demais na publicação correspondente, indicando que ele gera dados que concentram mais as chances de sorteio em poucos tópicos. Então quando você usa um dado como esse para gerar uma publicação, ela vai ficar mais concentrada em poucos tópicos, naqueles tópicos que têm mais chances de sair, os outros tópicos que tem chances muito pequenas, praticamente não serão sorteados. Observa-se pelos valores das relevâncias tópicas na Tabela 3.1 que há textos em que apenas um tópico prevalece fortemente, como é o caso de *Enriching_Annotating.1*, cujo tópico prevalente é o 6 segundo a análise com o LDA.



Texto Descritivo:
Enriching_Annotating.1

annotating refers to the activity of making information about a digital object explicit by adding notes, metadata, keywords, tags or links to a digitized representation or to an annotation file associated with it. this can be in the form of explanatory annotations that comments or contextualize a passage, annotations that make structural or linguistic information explicit, as linked open data making the relationships between objects machine readable, or in the case of general metadata, adding information about the object as a whole.

Vantagens do LDA:

1. *Mais usado*: Portanto mais testado e por isso mais explorado (BLEI, 2012; MCCALLUM; WANG; CORRADA-EMMANUEL, 2007; VAYANSKY; KUMAR, 2020);
2. *Mais adaptável*: Ao uso com outros algoritmos de inferência (VAYANSKY; KUMAR, 2020);
3. *Mais aplicável*: E adaptável para uma série de dados diferentes, gerando uma classificação suave (GRIFFITHS; STEYVERS, 2004);
4. *Mais referenciado*: por ser o mais citado e referenciar outros modelos, existe uma grande quantidade de trabalhos anteriores que servem para análise, revisão e que podem ser ajustáveis para muitas tarefas (GRIFFITHS; STEYVERS, 2004; SHEN; SUN; SHEN, 2008; VAYANSKY; KUMAR, 2020; WANG; MCCALLUM, 2006);
5. *Mais generalista*: Maior usabilidade em diferentes contextos tais como: (a) dados de mídias sociais (HONG; DAVISON, 2010); (b)padrões da web (MCCALLUM; WANG; CORRADA-EMMANUEL, 2007); (c) semântica (ALGHAMDI; ALFALQI, 2015); (d) linguística (ZUO et al., 2016); (e) avaliação de documentos (BERGHOLZ et al., 2008); (f) padrões de redes (VAYANSKY; KUMAR, 2020); (g) modelos preditivos (BLEI; MCAULIFFE, 2010); (h) processamento de imagens (ZHOU; ZHOU; ZHANG, 2016); (i) processamento de vídeos (HOSPEDALES; GONG; XIANG, 2012); (j) comportamento (BLEI; GRIFFITHS; JORDAN, 2010); (k) análise de sentimentos (BAO et al., 2009) e (l) diversos outros assuntos ou tópicos (BLEI, 2012).
6. *Mais diversificado*: Possui diversas versões e variações para diversos contextos (ALGHAMDI; ALFALQI, 2015).
    i) Versões: (a) supervisionada (BLEI; MCAULIFFE, 2010), (b) semi supervisionada (WATANABE, 2020), (c) não supervisionada (BLEI; NG; JORDAN, 2003);
    11) Variações: Tais como (a) *Collective LDA – C-LDA* (SHEN; SUN; SHEN, 2008); (b) *Latent Dirichlet Mixture Model – LDMM* (CHIENA; LEEA; TANB, 2017); (c) *Matrix Factorization Through LDA – fLDA* (AGARWAL; CHEN, 2010); (d) *Hierarchical Latent Dirichlet Allocation – hLDA* (BLEI; GRIFFITHS; JORDAN, 2010); dentre outras.

Desvantagens:

1. *Dispersão em Big Data*: Em documentos muito grandes podem existir palavras que não são encontradas no conjunto de treino e teste, o que dificulta um bom resultado final (BLEI; NG; JORDAN, 2003; VAYANSKY; KUMAR, 2020; VORONTSOV; POTAPENKO, 2014);
2. *Falta de correlação entre documentos*: Segundo (BLEI; LAFFERTY, 2006a), no mundo real documentos e palavras são correlacionados (BLEI, 2012; BLEI; NG; JORDAN, 2003; VAYANSKY; KUMAR, 2020).

3.3 Atribuição de Tópicos por Correlação das Relevâncias

Como o próprio nome já sugere, o diferencial do modelo de Correlação das Relevâncias dos Tópicos, identificado por *CTM*, acrônimo em inglês para *Correlated Topic Model* (BLEI; LAFFERTY, 2007; BLEI; LAFFERTY, 2006b), é a inclusão de correlações existentes entre os tópicos na determinação das relevâncias tópicas. A correlação entre dois tópicos,

digamos A eB, é uma medida, em uma escala entre -1 e 1, que indica o comportamento relativo das respectivas relevâncias tópicas nos diferentes textos analisados. A extremidade negativa **-1 (menos um)** indica que os tópicos A e B são inversamente correlacionados, ou seja, o tópico A é relevante nos textos em que tópico B não o é, e vice-versa. A quantificação da ideia de um tópico ser relevante em um texto é feita observando a variação do valor da sua relevância tópica com respeito à média das relevâncias tópicas desse tópico tomada sobre todos os textos. Nesse sentido, o caso extremo da correlação de valor -1 ocorre se a diferença da relevância tópica de A da média varia na mesma proporção que o tópico B dista da sua média, com sinal invertido.

Por outro lado, a extremidade positiva **1 (um)** reflete a situação em que os tópicos são igualmente relevantes, ou seja, o tópico A é relevante em um texto se e somente se o tópico B é relevante no mesmo texto. Entre esses dois extremos, a centralidade **0 (zero)** indica que não há nenhuma correlação entre os tópicos A e B no sentido que a relevância de um em um texto não traz nenhum indício da relevância tópica a respeito do outro no mesmo texto. No entanto, deve ser ressaltado que a correlação não é uma forma de distância semântica entre tópicos, mas apenas uma medida sobre quanto os tópicos tendem a aparecer, ou não, concomitantemente nos textos.

No modelo CTM, a sinalização das correlações existentes entre os tópicos é fetia através de uma matriz de covariâncias e da adoção de uma distribuição que leva em conta essa matriz na determinação das relevâncias tópicas de cada texto.

Vantagens do CTM:

1. *Mostra o relacionamento entre os diferentes tópicos*: os casos de corpus cujos textos possuem correlações entre si ocorrem muito frequentemente (VAYANSKY; KUMAR, 2020), argumento este apoiado pelos próprios autores do CTM (BLEI; LAFFERTY, 2007; BLEI; LAFFERTY, 2006b), conforme relato em (CAO et al., 2009).

2. *Visualização de dados*: segundo (BLEI; LAFFERTY, 2007), a determinação de correlações entre tópicos torna possível realizar diversas análises adicionais, incluindo visualizações gráficas. (VAYANSKY; KUMAR, 2020) explicam que esse artifício pode expor os dados de forma mais fácil e organizada para o usuário, possibilitando a flexibilização de tarefas, a exploração de grandes quantidades de documentos e a execução de tarefas colaborativas entre os membros da equipe de pesquisa.

Desvantagem:

Complexidade de uso: o ponto de atenção do CTM está relacionado à complexidade das contas matemáticas envolvidas, ou seja, caímos aqui na necessidade de maior poder computacional (KHALIFA et al., 2013) e maior tempo de processamento (ZHAO et al., 2015) nos cálculos dos métodos de resolução.

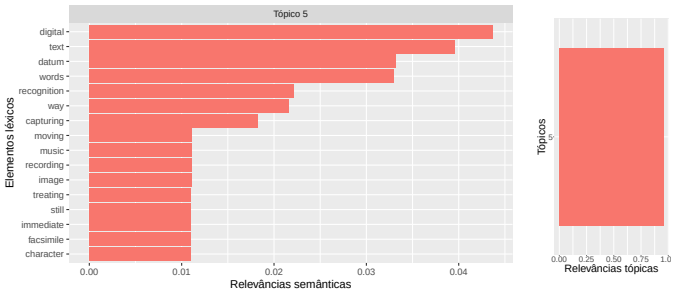
3.4 Atribuição de Tópicos por Estrutura

As características que diferem o modelo de atribuição de tópicos por estrutura (*Structural Topic Model – STM*) (ROBERTS et al., 2013; ROBERTS; STEWART; TINGLEY, 2019) são, além das correlações entre textos como no CTM, o agrupamento dos textos para construção de relevâncias tópicas e semânticas diversificadas e uma estratégia de inferência que fornece uma abordagem computacionalmente eficiente para o ajuste do modelo que é

suficientemente eficiente para a análise prática de uma quantidade expressiva de textos.O modelo combina características de 3 outros modelos já existentes (BLEI; LAFFERTY, 2007; EISENSTEIN; AHMED; XING, 2011; MIMNO; MCCALLUM, 2012).

A fim de estabelecer perfis específicos a grupos de textos, o método STM leva em consideração informações associadas aos textos na forma de metadados, condicionando antecipadamente a análise de acordo com conhecimentos disponíveis sobre os textos (BLEI; LAFFERTY, 2007; ROBERTS et al., 2013; ROBERTS et al., 2014; ROBERTS; STEWART; TINGLEY, 2019). O agrupamento dos textos é definido através de valores de metadados, sendo um grupo definido pelos textos que compartilham valores comuns desses metadados. Exemplos de metadados que podem ser utilizados para a definição de grupos na análise de publicações são o estrato da *TaDiRAH*, o tipo, a data, o local ou o veículo de publicação, dentre outros. Posto de maneira ilustrativa, o conjunto de textos de cada grupo é visto como uma amostra de textos produzidos por um mesmo autor exclusivo, distinto dos autores dos demais grupos. Dessa forma, busca-se modelar o comportamento de um único autor fictício dos textos por grupo segundo a descrição de (ALGHAMDI; ALFALQI, 2015) sobre o processo de "imitação" da escrita humana feita pelos modelos generativos probabilísticos.

O método de análise dos tópicos identifica perfis de diferentes fontes como resultado de um agrupamento de acordo com o estrato a que está atribuído cada conjunto de publicações. Ressalte-se que no método LDA, seguindo pela mesma ilustração, todos os textos, mesmo de grupos diferentes, seriam gerados pela mesma pessoa fictícia, ou seja, pelo único perfil generativo das relevâncias tópicas. A introdução de metadados e o agrupamento deles decorrente permite também estabelecer que as correlações entre os tópicos sejam ponderadas segundo grupos de textos. Essa propriedade adicional do método STM lhe confere a possibilidade de identificação de a quais dentre os grupos o perfil de tópicos estaria mais inclinado. Finalmente, ainda há a possibilidade de utilização de um conjunto de relevâncias semânticas distinto por grupo de textos, o que torna possível apresentá-los sob diferentes abordagens ou pontos de vista (ALGHAMDI; ALFALQI, 2015). A ilustração abaixo mostra o efeito de agrupamento no exemplo da análise de *Capturing-Creating_Datarecognition.1* comparado à análise anterior com o método LDA.



Texto Descritivo:
Capturing-Creating_Datarecognition.1

data recognition, e.g. optical character recognition (ocr), refers to the process of treating the immediate products of digital data capturing (recording or imaging), such as digital facsimiles of texts or of music score, in a way to extract discrete, machine readable units from them, such as plain text words, musical notes, still or moving image elements.

Existem outros modelos que usam abordagem similares de inserir informações diretamente no corpus para alterar os dados de tópicos e os dados de elementos léxicos, tais como informações de tempo (BLEI; LAFFERTY, 2006a), autor (ROSEN-ZVI et al., 2012), ideologia (AHMED; XING, 2010) e lugar (EISENSTEIN et al., 2010), mas o STM se distingue deles porque (1) eles são específicos demais, (2) no STM essas informações são inseridas diretamente no modelo e por isso (3) o STM é adaptável a diversas realidades distintas (ROBERTS et al., 2013; ROBERTS; STEWART; TINGLEY, 2019).

Vantagens do STM:

1. *Pacote R para implementação*: análises variadas podem ser levadas a cabo sem a necessidade de usar modelos específicos ou desenvolver novos modelos específicos a cada caso, tornando-se adequado a estudos experimentais (ROBERTS et al., 2013);
2. *Flexibilidade*: a capacidade de lidar com agrupamentos torna a análise de respostas abertas mais fácil e mais reveladora (ROBERTS et al., 2013; ROBERTS et al., 2014);
3. *Generalidade*: no caso de não haver informações associadas aos documentos, o modelo faz uma correlação de tópicos como no modelo CTM de (BLEI; LAFFERTY, 2006b).
4. *Usabilidade*: o método de resolução não exige maior poder computacional (KHALIFA et al., 2013; VORONTSOV; POTAPENKO, 2014) e nem maior tempo de processamento (ZHAO et al., 2015) comparativamente a outros modelos generativos probabilísticos semelhantes;
5. *Adaptabilidade*: do modelo a muitas variedades de propósitos (ROBERTS et al., 2014).

Desvantagem:

Pouco referenciado: Existem poucos estudos que usam o STM, e por isso temos poucos exemplos práticos de estudos para verificar e comparar com outros modelos.

3.5 Classificação Supervisionada

Os modelos de aprendizado supervisionado de análise textual têm como pressuposto a existência de exemplares rótulados manualmente. Esses textos rotulados, quando expostos a um algoritmo de aprendizado supervisionado, são analisados a fim de determinar padrões de combinação de elementos léxicos presentes no texto, associando-os aos respectivos rótulos. Uma vez determinadas essas associações, o algoritmo pode buscar e encontrar padrões semelhantes em outros textos não rotulados manualmente, e assim receberem os mesmos rótulos destes. Modelos de classificação supervisionada são facilmente adaptáveis às realidades diversas de acordo com o contexto de cada demanda. Existem muitos modelos e variáveis de modelos, por isso, escolher um vai depender de cada contexto (BLEI; MCAULIFFE, 2010; BREIMAN, 2001; CORTES; VAPNIK, 1995; VIKRAMKUMAR; B; TRILOCHAN, 2014; WATANABE, 2020).

Em síntese, a classificação supervisionada por meio de análise por tópicos segue uma sequência bem definida de passos. Primeiramente, é preciso reunir um conjunto de textos já rotulados, ou "classificados" no jargão normalmente utilizado, por especialistas. No caso da classificação de publicações, o corpus empregado nesta dissertação é formado por um conjunto de publicações classificadas segundo os estratos da *TaDiRAH* na plataforma

Zotero (DARIAH-UE CONSORTIUM, 2012). Adicionalmente, um conjunto de publicações ainda não classificadas constitui os dados alvo de uma aplicação prática. No caso dos experimentos de avaliação que são descritos em capítulos posteriores, são usadas publicações já classificadas com o intuito de verificar em que medida os métodos propostos produzem resultados coincidentes com a classificação produzida por especialistas. Nesses experimentos, os resultados obtidos com o método *sLDA – Supervised Latent Dirichlet Allocation* são usados como referência de comparação com os métodos propostos nos próximos capítulos (BLEI; MCAULIFFE, 2010).

O sLDA é um classificador de textos segundo a abordagem de tópicos que pode ser usado para organizar, compreender e resumir grandes quantidades de textos, uma vez que podemos rotular os documentos com nomes de classes que não tem necessariamente a ver com os textos, ou os tópicos, mas sim com o tipo de nome de classes que queremos dar a eles. O *sLDA* aprende o nome da classe no contexto da abordagem de tópicos. Ele descobre os temas ocultos e associa esses temas às classes pré definidas pelo usuário, por isso é muito útil em atender a demandas distintas relacionadas às informações presentes nos textos. Partindo do modelo geral de análise léxica por tópicos descrito neste Capítulo, as relevâncias tópicas dos textos são usadas para estabelecer os padrões associados aos rótulos da classificação. Mais especificamente, o padrão das publicações de um rótulo é definido por um conjunto de coeficientes de uma regressão logística de vetores representativos dos textos. Esse vetor representativo de um texto, digamos $\mathbb{d}$, contém, para cada tópico t, a quantidade de vezes em que as relevâncias semânticas de t são utilizadas para a geração de elementos léxicos de $\mathbf{d}$. O sLDA produz como saída o conjunto de coeficientes de cada rótulo. A etapa de classificação determina as quantidades de utilização dos tópicos para cada texto alvo e, com a ajuda dos coeficientes da regressão logísitca, determina o rótulo que cada texto alvo deve receber.

De acordo com o objetivo deste trabalho, cada rótulo na classificação com sLDA representa um estrato da *TaDiRAH* a fim de que as publicações sejam classificadas de acordo com a taxonomia. Porém, o fato de o sLDA empregar uma regressão logística pressupõe, por definição, a inexistência de sobreposição. Já vimos no Capítulo 2 que existem muitas sobreposições nas nossas publicações classificadas manualmente (DARIAH-UE CONSORTIUM, 2012), isso graças à natureza da taxonomia e da interdisciplinaridade presente nas Humanidades Digitais. Por essa razão, propomos outro método no Capítulo 4, além de utilizarmos o próprio método *LDA* supervisionado de uma forma específica.

4 CLASSIFICAÇÃO POR TÓPICOS

Neste capítulo, descrevemos os métodos de classificação que propomos para lidar com uma das questões que envolvem classificar textos segundo a *TaDiRAH*: As sobreposições de documentos classificados em diferentes estratos da taxonomia. Trataremos a questão do desbalanceamento no capítulo seguinte.

No contexto de publicações em Humanidades Digitais, usamos o mapeamento prévio de elementos léxicos feito pela taxonomia TaDiRAH, juntamente com os textos classificados manualmente por especialistas de Humanidades Digitais na análise léxica por tópicos buscando por padrões regulares sintáticos e semânticos a partir da análise de frequências de ocorrências concomitantes de elementos léxicos, já que estes elementos léxicos são os compostos das estruturas sintáticas e semânticas (DEERWESTER et al., 1990).

Começamos com as preliminares sobre o nosso contexto de classificação por tópicos, os conceitos básicos que norteiam os métodos de classificação que propomos, seguida de suas descrições: trata-se de três métodos supervisionados propostos a partir de análises por tópicos no cenário de sobreposições. Os dois primeiros são métodos bayesianos inspirados no método não supervisionado de YAU et al. (2014), diferenciando-se pelo método de análise por tópicos aplicado: O *Latent Dirichlet Allocation - LDA* (BLEI; NG; JORDAN, 2003) e o *Structural Topic Model - STM* (ROBERTS et al., 2013).

Terminamos o capítulo abordando o terceiro método, a classificação com LDA supervisionado - sLDA (BLEI; MCAULIFFE, 2010), e as atitudes tomadas com o objetivo de contornar os problemas ocasionados pelas sobreposições de documentos em mais de um estrato da *TaDiRAH*, pois como já explicamos no Capítulo 3 o classificador LDA supervisionado, pressupõe que não haja sobreposições de estratos nas publicações devido ao método utilizado para gerar o dado de atribuição de estrato.

4.1 Preliminares

Como já mencionado no Capítulo 2, nas publicações classificadas manualmente por especialistas de Humanidades Digitais retiradas de DARIAH-UE CONSORTIUM (2012), encontramos muitas sobreposições de classificação nos estratos da *TaDiRAH*. Dentre as causas descritas, abordamos as seguintes:

1. A interdisciplinaridade do campo das Humanidades Digitais.
2. A dinâmica própria das atividades de pesquisa em Ciências Humanas.
3. O imbricamento das atividades especificadas nos elementos léxicos que definem os estratos superiores da taxonomia *TaDiRAH*.
4. A forma de uso e definições da própria taxonomia.

Como consequência disso, temos que: para classificar publicações de Humanidades Digitais precisaremos necessariamente lidar com a questão das sobreposições. Ou seja, encontrar mecanismos para amenizar os seus efeitos sobre os resultados da classificação.

O conjunto de métodos que tratamos neste capítulo é uma abordagem inspirada no método

de classificação não-supervisionada assistida descrita em YAU et al. (2014) A abordagem proposta em YAU et al. (2014) pode ser resumida da seguinte forma. Trata-se de um método de classificação assistida de publicações científicas baseada em análise por tópicos em três etapas:

1. Aplicação de método de análise por tópicos para obter as relevâncias tópicas. A quantidade de tópicos T usada nessa análise é algumas vezes superior à quantidade de estratos.

2. Análise artesanal dos elementos léxicos mais relevantes de cada tópico a fim de interpretar o seu sentido e, com isso, atribuí-lo aos estratos. Ao final desta etapa, cada tópico está associado a um estrato, com cada estrato sendo potencialmente representado por múltiplos tópicos.

3. Cálculo de relevância de cada estrato em cada publicação a partir da relevância dos seus tópicos determinada pela análise por tópicos.

A atribuição de cada publicação $\mathbf{d}_i$ a um estrato é feita a partir das relevâncias tópicas, em linhas gerais, da seguinte forma. Primeiramente, os tópicos com relevância tópica em $\mathbf{d}_i$ são verificados. Em seguida, as relevâncias tópicas dos tópicos associados a um mesmo estrato são somadas, obtendo a relevância de cada estrato em $\mathbf{d}_i$. Finalmente, $\mathbf{d}_i$ é classificado como pertencente ao estrato com maior relevância.

Em um exemplo hipotético com intuito meramente ilustrativo, suponhamos que no caso das relevâncias tópicas da Tabela 3.1, os tópicos sejam associados estratos denominados A, B e C, com o tópico 3 associado ao estrato A, tópicos 1 e 5 associados ao estrato B e os demais tópicos associados ao estrato C. Nesse cenário, o documento *Analyzing_Spatialanalysis.1* recebe uma influência de 51,86% do estrato A, 47,65% do estrato B e 0,49% do estrato C. Nesse cenário hipotético, o documento *Analyzing_Spatialanalysis.1* seria associado ao estrato A.

Retomando esse mesmo movimento com todos os demais documentos do corpus, cada um deles estará associado a um estrato apenas.

Nos experimentos descritos a seguir apresentamos uma versão supervisionada voltada para classificação de textos com sobreposição de classes. Seguimos a abordagem de YAU et al. (2014). A fim de lidar com essas sobreposições, uma abordagem bayesiana substitui a regressão empregada nos classificadores baseados no LDA de que tratamos no Capítulo 5.

No entanto, adotamos uma modificação com relação ao método descrito em YAU et al. (2014) no que se refere à atribuição dos tópicos aos estratos. A fim de evitar a etapa artesanal do método original, adotamos um procedimento automatizado:

1. Análise por tópico em todas as publicações que já estão classificadas e;

2. Associação automatizada de cada tópico a um estrato.

Cada método de análise de tópicos tem seu próprio procedimento, como descrito nas seções a seguir.

4.2 Pertinência Tópica das Publicações e dos Textos Descritivos da *TaDiRAH*

Com o intuito de reforçar o perfil de cada um dos estratos, adotamos o procedimento de juntar os textos descritivos das subcategorias da *TaDiRAH* abordados no Capítulo 2 ao conjunto de publicações em cada estrato, cada descrição dentro da sua subcategoria equivalente. Sendo assim, o corpus $\mathbf{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_D\}$ passa a ser formado pelas publicações classificadas segundo a *TaDiRAH* e pelos textos descritivos das diversas subcategorias dos estratos $\mathbf{E} = \{Analyzing, Capturing - Creating, Disseminating - Storing, Enriching, Interpreting\}$. Esse corpus estendido tem um perfil de sobreposição de estratos modificados quanto ao que é analisado no Capítulo 2Para cada estrato $e \in \mathbf{E}$a notação $\mathbf{D}_e$ os documentos do corpus $\mathbf{D}$ classificados previamente no estrato $e \in \mathbf{E}$. Portanto, $\mathbf{D}_e = \{\mathbf{d}_i \in \mathbf{D} \mid \mathbf{d}_i$ pertence ao estrato $e\}$, com $|\mathbf{D}_e| = D_e$.

De acordo com a Figura 4.1, sobreposições de estratos no corpus $\mathbf{D}$ acontecem, e isso devido a múltiplas ocorrências de algumas publicações em $\mathbf{D}$. Uma dupla ocorrência corresponde à existência de dois documentos idênticos identificados separadamente em $\mathbf{D}$, digamos $\mathbf{d}_i$ e $\mathbf{d}_j$, tais que $i \neq j$ e $\mathbf{d}_i = \mathbf{d}_j$. Em tal situação, as duas ocorrências estão associadas a estratos distintos. Mais precisamente, $\mathbf{d}_i \in \mathbf{D}_e$ e $\mathbf{d}_j \in \mathbf{D}_{e'}$, sendo $e \neq e'$. Múltiplas ocorrências são generalizações de duplas ocorrências para dois ou mais documentos. A forma de apresentação utilizada na Figura 4.1 consite na identificação das 32 possíveis interseções das cinco formas nomeadas segundo os estratos da *TaDiRAH*. As distintas interseções, incluindo aquelas de uma só forma, são identificadas com cores distintas. As interseções que envolvem mais de uma forma indicam possíveis sobreposições de documentos nos estratos respectivos. Os números são indicativos das quantidades de documentos que ocupam as interseções de formas em que aparecem. Com essa apresentação, podemos observar como os 261 documentos do estrato *Analyzing* estão distribuídos nas interseções entre formas.

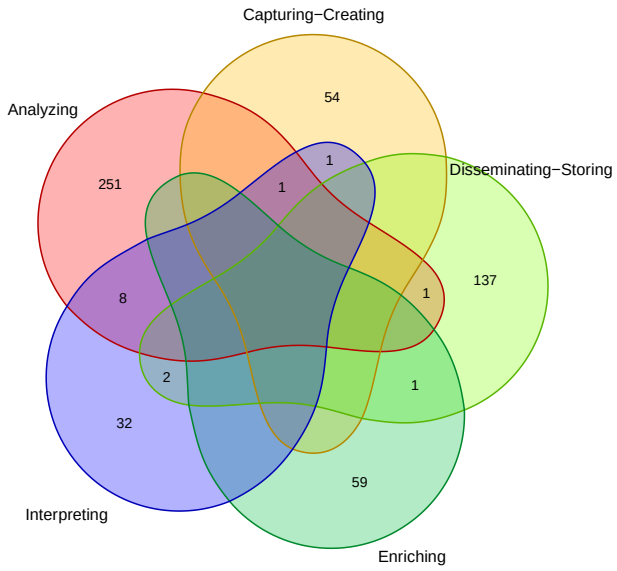


Figura 4.1: Quantidades de sobreposições entre os estratos, incluindo publicações e textos descritivos.

Na sequência da seção, supomos que tenhamos os tópicos $\mathbf{T}$ e as respectivas relevâncias tópicas resultantes de uma análise por tópicos de todo o corpus estendido $\mathbf{D}$. Com as

relevâncias tópicas disponíveis, passamos a associar cada tópico obtido na análise aos estratos da *TaDiRAH*. Passamos da abordagem anterior onde para cada tópico era associado a um estrato por inspeção humana, para uma distribuição de probabilidades onde determinamos uma medida de aderência de cada tópico em cada estrato. A seção é então finalizada com uma descrição de forma de aplicação do método em combinação com o LDA e com o STM.

4.2.1 Definição do Método

O método que propomos consiste em um desenvolvimento estatístico básico para obter $P(e \mid t_j)$, sendo e um estrato e t_j um tópico. Nesse sentido, a primeira providência é identificar a relevância de cada tópico em cada estrato, ou *prevalência tópica*, que seria a medida que expressa a proporção de ocorrência do tópico $t_j \in \mathbf{T}$ dentro do estrato $e \in \mathbf{E}$ através da probabilidade posterior

$$\begin{aligned} P(t_j \mid e) &= \sum_{\mathbf{d}_i \in \mathbf{D}_e} P(t_j \mid \mathbf{d}_i) P(\mathbf{d}_i \mid e) \\ &= \frac{1}{P(e)} \sum_{\mathbf{d}_i \in \mathbf{D}_e} P(t_j \mid \mathbf{d}_i) P(e \mid \mathbf{d}_i) P(\mathbf{d}_i). \end{aligned}$$

O cálculo da prevalência tópica $P(t_j \mid e)$ de e em t_j decorre das seguintes estimativas.

- A probabilidade de ocorrência de cada estrato $e \in \mathbf{E}$, $P(e)$, que expressa a razão entre a quantidade de publicações em e e a quantidade total de publicações no corpus, dada por $P(e) = \frac{D_e}{D}$.
- Análise por tópicos prévia de $\mathbf{D}$ determina $P(t_j \mid \mathbf{d}_i)$ para todos $t_j \in \mathbf{T}$ e $\mathbf{d}_i \in \mathbf{D}$, o que corresponde à probabilidade de ocorrer cada tópico t_j condicionada por cada publicação $\mathbf{d}_i$.
- A probabilidade de ocorrência do estrato e condicionada por cada publicação $\mathbf{d}_i \in \mathbf{D}$ é dada por

  $$P(e \mid \mathbf{d}_i) = \begin{cases} 0, & \text{se } e \notin \mathbf{E_D}(\mathbf{d}_i) \\ \frac{1}{|\mathbf{E_D}(\mathbf{d}_i)|}, & \text{caso contrário} \end{cases}$$

  onde $\mathbf{E_D}(\mathbf{d}_i) = \{e \in \mathbf{E} \mid \mathbf{d}_i \in \mathbf{D}_e\}$ é o conjunto de estratos aos quais o documento $\mathbf{d}_i$ é atribuído na *TaDiRAH*. Observe que essa medida significa distribuir igualmente as chances de $\mathbf{d}_i$ aparecer em cada um dos estratos. se $\mathbf{d}_i$ aparecer em um estrato apenas, digamos e, então tem-se $|\mathbf{E_D}(\mathbf{d}_i)| = 1$ e $P(e) = 1$ (o que corresponde a 100%) e, para qualquer outro estrato $e' \neq e$, $P(e') = 0$. De forma similar, se $\mathbf{d}_i$ pertencer a dois estratos e_1 e e_2, então $|\mathbf{E_D}(\mathbf{d}_i)| = 2$, $P(e_1) = P(e_2) = 0,5$ e a probalidade é nula para os demais estratos. O mesmo comportamento se repete para os casos em que $|\mathbf{E_D}(\mathbf{d}_i)|$ tem valores maiores que 2.
- A probabilidade de ocorrência de $\mathbf{d}_i$ é também dividida igualmente por todos os documentos. Portanto, $P(\mathbf{d}_i) = \frac{1}{D}$.

Finalmente chegamos à definição da *pertinência tópica*, que é a expressão da verossimilhança $P(e \mid t_j)$ do tópico t_j relativo ao estrato e.

Usando o Teorema de Bayes, obtemos

$$\begin{aligned} P(e \mid t_j) &= \frac{P(t_j|e)P(e)}{P(t_j)} \\ &= \frac{\sum_{\mathbf{d}_i \in \mathbf{D}_e} P(t_j|\mathbf{d}_i)P(e|\mathbf{d}_i)P(\mathbf{d}_i)}{\sum_{\mathbf{d}_i \in \mathbf{D}} P(t_j|\mathbf{d}_i)P(\mathbf{d}_i)} \\ &= \frac{\sum_{\mathbf{d}_i \in \mathbf{D}_e} P(t_j|\mathbf{d}_i)P(e|\mathbf{d}_i)}{\sum_{\mathbf{d}_i \in \mathbf{D}} P(t_j|\mathbf{d}_i)} \end{aligned}$$

A premissa intrínseca à análise acima descrita é que a prevalência e a pertinência tópica guardem uma correlação com as respectivas características semânticas das publicações e dos estratos. A partir daqui veremos como que a pertinência tópica se passa com cada um dos métodos abaixo.

4.2.2 Aplicação com LDA

Introduzimos a análise da *pertinência tópica* no LDA definindo a quantidade de tópicos através de procedimentos estatísticos, determinado em 200, levando-se em conta o método de definição da quantidade de tópicos em <https://juliasilge.com/blog/evaluating-stm/>. No LDA existem ferramentas para realizar essa análise (GROSSETTI, 2021; GROSSETTI; LEWIS, 2019; LEWIS; GROSSETTI, 2019).

Na Figura 4.2, são exibidos os cinco tópicos mais pertinentes em cada um dos 5 estratos da *TaDiRAH* do nosso trabalho (*Analyzing, Enriching, Capturing-Creating, Disseminating-Storing, Interpreting*) identificados por cores distintas e pelos elementos léxicos mais relevantes em cada tópico, devidamente identificados.

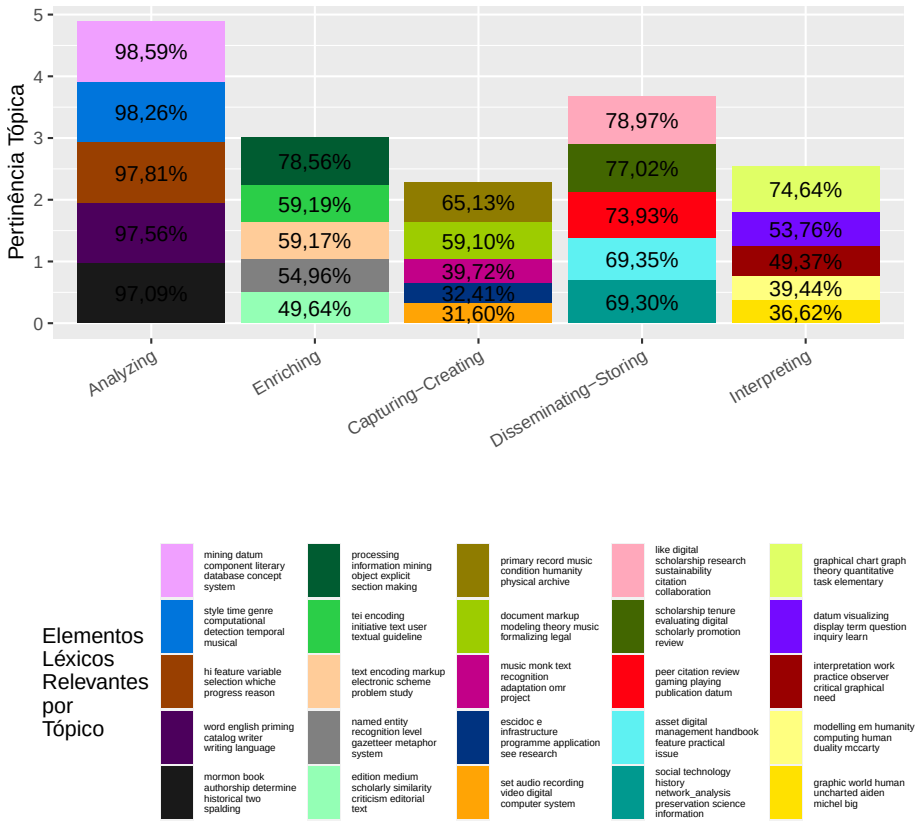


Figura 4.2: Cinco maiores pertinências tópicas por estrato e elementos léxicos mais relevantes por tópico segundo LDA.

Passamos agora a análise da pertinência tópica dos cinco tópicos principais em cada estrato com a finalidade de esclarecer pontos relevantes em cada grupo de publicação com o filtro do LDA.

Para facilitar a leitura dividimos os estratos em dois grupos:

- *1º Grupo*: os estratos cujos tópicos associados a eles são de pertinência tópica alta,

acima de 90%. Cada um desses tópicos prevalece praticamente exclusivamente no respectivo estrato e uma pequeníssima parte deles está sendo dividido com os demais estratos. Com uma simples verificação dos elementos léxicos que identificam o tópico e uma visão geral dos substratos, podemos rapidamente identificar de que se trata cada tópico dentro dos temas da *TaDiRAH*. Vejamos alguns exemplos:

  - *Analyzing*: percebemos a identificação dos tópicos com o segundo nível das subcategorias de *Analyzing* em (DHTAXONOMY, 2020): *Content Analysis*, *Network Analysis*, *Relational Analysis*, *Spatial Analysis*, *Structural Analysis*, *Stylistic Analysis* e *Visual Analysis*. Segundo observado no Capítulo 2, os temas, elementos de análise dos tópicos, são mais nítidos nos substratos de primeiro e segundo nível, ficando cada vez mais misturados (menos nítidos em um único estrato) do terceiro nível de subcategorias em diante. Como exemplo, citamos o tópicos identificados pelos respectivos elementos léxicos {*mining datum, component literary, database concept, system*} e {*style time genre, computational, detection temporal, musical*}. Tais tópicos tratam de análise de métodos e, provavelmente também, de conteúdos.
  - *Disseminating-Storing*: os tópicos pertinentes associam-se aos temas dos substratos de primeiro nível *Collaborating*, *Commenting*, *Communicating*, *Crowdsourcing*, *Publishing*, *Sharing* e *Teaching*. Como exemplo citamos o tópico {*like digital, scholarship research, sustainability, citation, collaboration*}, claramente relacionado a *Publishing* e *Teaching*.

- *2º Grupo*: formado pelos estratos cujos os tópicos associados a eles são de pertinência tópica média (entre 60% e 90%) e baixa (abaixo de 60%).
  - *Enriching*: os tópicos pertinentes têm dificuldade de serem associados aos temas da *TaDiRAH*, ou seja, não se traduzem claramente para nós, devido a sua baixa pertinência tópica, com exceção de um único tópico com pertinência alta que associamos a *Annotating* e *Editing*.
  - *Capturing-Creating*: aqui temos um único tópico com pertinência alta que associamos aos temas *Converting*, *Web Development* e *Programming*. O que chama a atenção neste estrato é que apesar da pertinência tópica ser média e baixa baixa, ainda assim conseguimos associar os tópicos aos temas da *TaDiRAH* em *Capturing* (*Converting*, *Data Recognition*, *Discovering*, *Extracting*, *Gathering*, *Imaging*, *Recording* e *Transcribing*) e *Creating* (*Designing*, *Programming*, *Translating*, *Web Development* e *Writing*). Uma explicação para esse fato pode estar relacionada à baixa quantidade de sobreposições entre esse estrato e os demais. Na Figura 4.2 vemos que as únicas duas publicações com sobreposição em *Capturing-Creating* é com *Analyzing*. Esse fato ajuda a preservar a temática do estrato segundo os estratos da *TaDiRAH*.
  - *Interpreting*: os temas da *TaDiRAH* (*Contextualizing*, *Modeling* e *Theorizing*) são mais difíceis de terem associação direta com os tópicos, o que diminui a pertinência tópica. Comparativamente com *Disseminating-Storing*, *Interpreting* é o menor estrato e consequentemente é o que tem mais dificuldade de ser encontrado, dado a essa dificuldade de associar os temas da *TaDiRAH* aos tópicos.

Sendo assim, o fato de haver muito mais textos classificados manualmente no estrato *Analyzing* e em *Disseminating-Storing* que nos demais estratos da *TaDiRAH*, conforme constatado na Figura 4.1, possibilita ao algoritmo aprender melhor os estratos que têm maior quantidade de publicações associadas. Os estratos *Analyzing* e *Disseminating-*

Storage são os que possuem a maior pertinência tópica, o que significa que as publicações muito identificadas com esses tópicos vão se identificar muito com esses estratos também. Por isso, a tendência é que as publicações desses estratos sejam bem identificadas pelo algoritmo, principalmente aquelas que estão muito influenciadas pelos cinco tópicos mais pertinentes analisados na Figura 4.2. Portanto, quanto maior a relevância tópica nas publicações de um estrato, maiores são as chances de as publicações desse estrato serem bem classificadas.

Pela Figura 4.2, observa-se que os estratos {*Analyzing*} são os que mais são identificados pelos tópicos reconhecidos pela análise.

Notamos também que ao contrário dos estratos cuja pertinência tópica ultrapassa os 90%, alguns estratos, como *Enriching* (principalmente) e *Interpreting*, cujas as cinco maiores pertinências tópicas, não são todas altas, revelam que a identificação das publicações nestes dois estratos tendem a ser menos precisa.

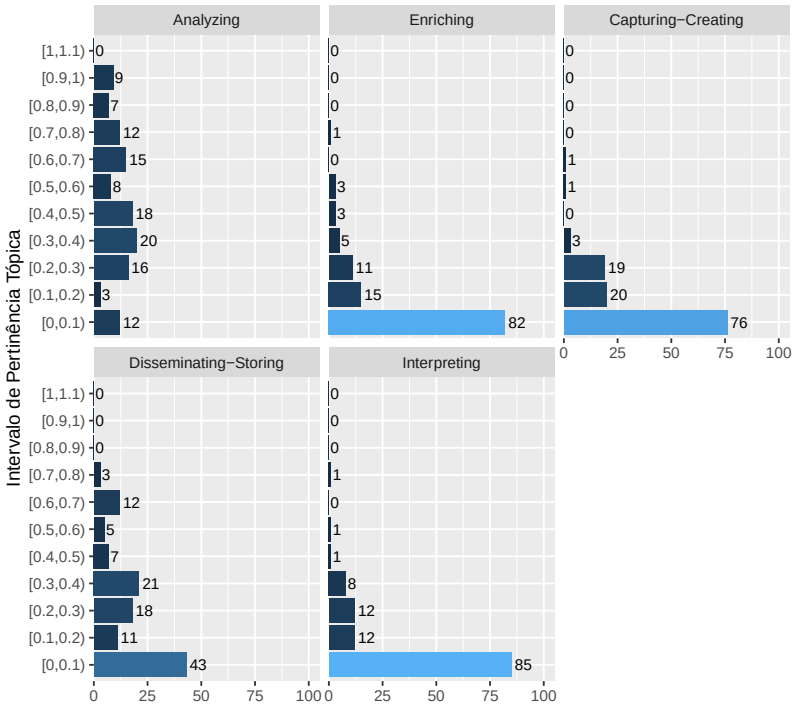


Figura 4.3: Quantidade de tópicos em cada intervalo de pertinência tópica por estrato segundo LDA.

Também podemos verificar que ter tópicos que não são tão pertinentes a cada estrato tem a ver naturalmente com o desbalanceamento. É uma consequência do desbalanceamento. Pois vemos que nos estratos que têm menos publicações, os seus tópicos perdem força, perdem representatividade, porque ficam pequenos, com relação aos outros. Os tópicos que aparecem nas publicações de *Enriching* não aparecem muito claramente nos tópicos que são identificados pela análise, por isso que eles não são tão pertinentes assim. Logo, vemos uma razão para que os estratos de *Analyzing* e *Disseminating-Storing* fiquem muito melhor classificados que os outros.

Passamos agora a Figura 4.3, onde observamos a quantidade de tópicos por intervalo de pertinência, sendo que o estrato *Analyzing* possui mais de 1/5 dos 200 tópicos da análise com pertinência tópica superior a 90%. Esses tópicos com pertinência tópica alta são

os que de fato aparecem e influenciam na análise. Isso demonstra uma tendência destes tópicos com pertinência alta arrastarem as publicações para este estrato. Assim sendo, um processo de classificação baseado na análise via LDA deve identificar bem as publicações que de fato pertencem a esse estrato. Além disso ainda possui bastante tópicos com pertinência tópica média, mostrando uma tendência a também classificar publicações de outros estratos.

Em *Enriching* já acontece o contrário, o maior grupo de tópicos encontrados no estrato está nos intervalos de pertinência da média para baixa, acontecendo algo semelhante em *Capturing-Creating*, sendo o pior caso, o de *Interpreting*, além disso eles ainda possuem poucas publicações. Como já vimos, poucas publicações produzem baixo número de pertinência tópica. O que significa dizer que se um tópico tem baixa pertinência em um estrato e alta pertinência em um outro, o documento onde o tópico se encontra mais relacionado tenderá a classificar a publicação daquele tópico.

Disseminating-Storing segue a tendência de *Analyzing* com boa concentração de alta pertinência tópica. Isso provoca uma boa associação das publicações onde estes tópicos são relevantes a este estrato. Desta forma podemos entender melhor os movimentos de associação das publicações aos estratos. E porque os estratos de *Analyzing* e *Disseminating-Storing* tendem a funcionar bem (boa classificação) e os outros tendem a ter dificuldades de serem identificados.

4.2.3 Aplicação com STM

Vejamos agora a análise da pertinência tópica nas publicações com o STM, onde poderemos perceber que os resultados são parecidos, porém o STM consegue melhorar um pouco a pertinência dos tópicos, e isso tende a equilibrar um pouco mais os resultados finais.

Antes porém, precisamos definir alguns parâmetros, sendo assim, definimos:

1. A quantidade de tópicos para análise STM fixada em 120;
2. Os grupos para correlação entre estratos definimos por:

- Tipo de publicação;
- Ano de publicação e;
- Estrato;

3. Também definimos os grupos para os dados de elementos léxicos por estrato.

Dito isso, sigamos para a análise da pertinência tópica dos cinco tópicos principais em cada estrato segundo o STM visualizada na Figura 4.4.

Podemos observar fenômenos parecidos com os já vistos na Análise anterior, em que *Analyzing* e *Disseminating-Storing* se impõem acima dos demais estratos com muitos tópicos apresentando pertinência tópica alta. O que equivale a dizer que esses tópicos possuem mais facilidade de serem relacionados com os temas da *TaDiRAH*, arrastando os documentos em que estes tópicos possuem mais relevância para serem classificados nestes estratos.

Os piores resultados continuam com *Enriching* e *Interpreting*, sendo este último pior que o anterior. Isto porque *Enriching* possui poucos tópicos com pertinências média e alta, enquanto que *Interpreting* não possui nenhum tópico com pertinência alta, possui

poucos com pertinência média e muitos com pertinência baixa. Dos menores estratos, *Capturing-Creating*, é privilegiado pela baixa sobreposição de publicações com os outros estratos, o que facilita para o algoritmo a identificação de elementos léxicos exclusivos, e a consequente identificação desses tópicos com os temas da *TaDiRAH* mesmo com pertinências tópicas entre média e baixa.

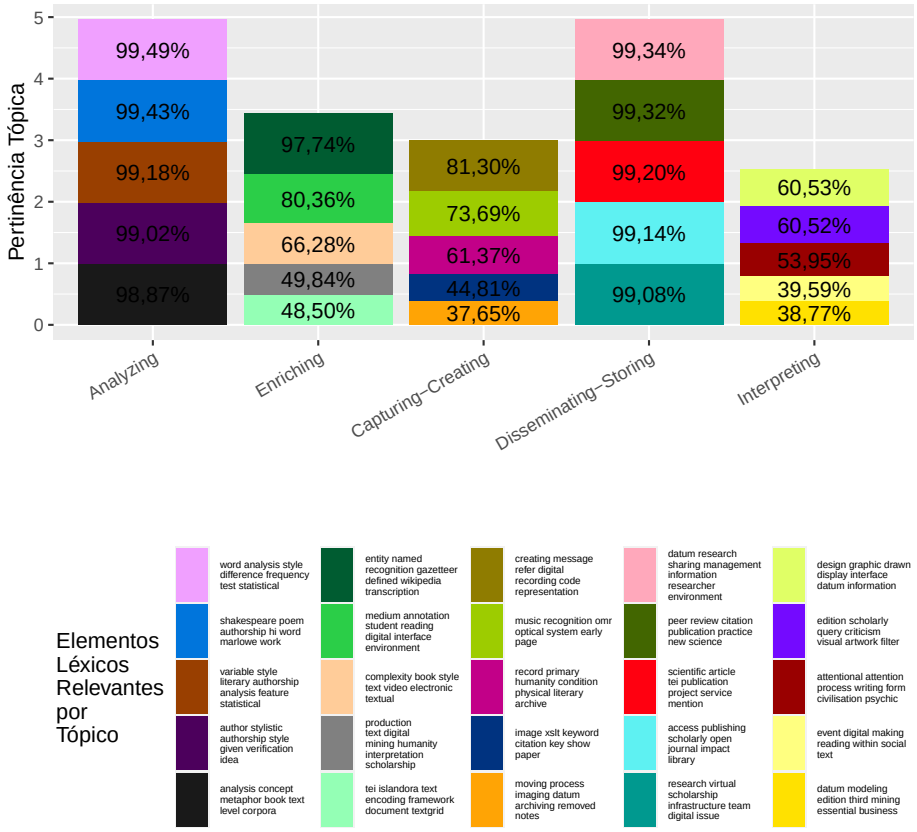


Figura 4.4: Cinco maiores pertinências tópicas por estrato e elementos léxicos mais relevantes por tópico segundo STM.

Voltamos a associar esse fenômeno ao desbalanceamento que provoca o fortalecimento dos estratos com muitas publicações associadas a eles em detrimento daqueles com poucas publicações. Entretanto, graças à correlação de metadados feita pelo STM, temos um pouco mais de facilidade de associar os grupos de elementos léxicos que identificam os tópicos aos temas da *TaDiRAH*, como veremos abaixo nos exemplos:

1. *Analyzing*: o tópico {*word analysis style, difference frequency, test statistical*} é visivelmente relacionado a *Stylistic Analysis*, *Relational Analysis* e *Structural Analysis*.

2. *Disseminating-Storing*: tópico {*datum research, sharing management, information, researcher, environment*} relacionado a *Publishing*.

3. *Enriching*: o tópico {*entity named, recognition gazetteer, defined wikipedia, transcription*} aparenta associação com *Annotating*.

4. *Capturing-Creating*: o tópico {*creating message, refer digital, recording code, representation*} mostra associação com reconhecimento ótico de música.

5. *Interpreting*: tópico {*design graphic drawn, display interface, datum information*} com associação com visualização gráfica.

Vemos uma vez mais que $\{Analyzing, Enriching\}$ são os estratos que mais são identificados pelos tópicos reconhecidos pela análise. Esse fato pode estar associado a quantidade de publicações previamente classificadas e, no caso do STM, também aos metadados e publicações relacionadas.

Quanto ao efeito do desbalanceamento, podemos notar que os estratos com mais amostras são mais determinantes sobre os tópicos tanto no LDA quanto no STM. O STM conseguiu melhorar um pouco os resultados da pertinência tópica, o que facilita a associação dos elementos léxicos dos tópicos com os temas da *TaDiRAH*.

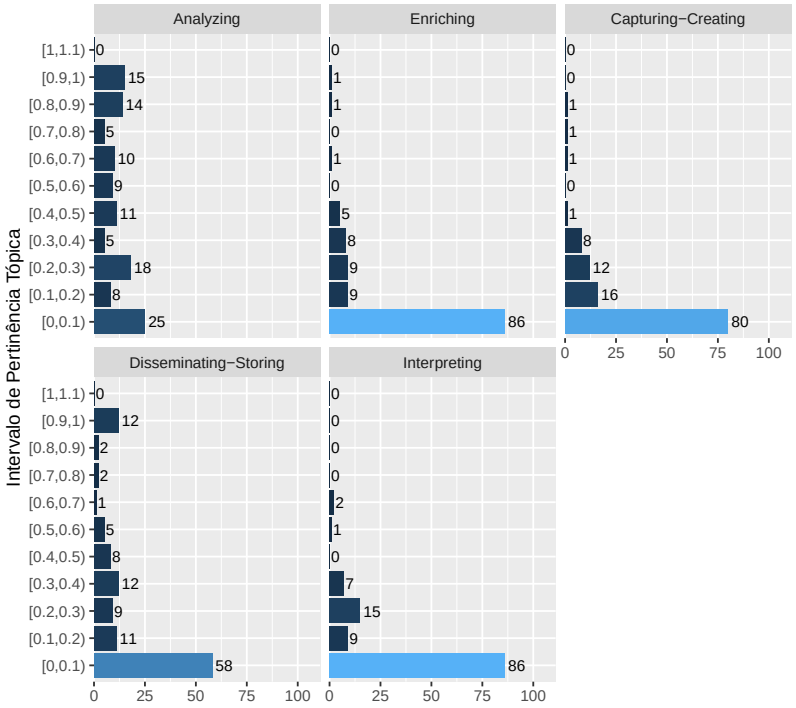


Figura 4.5: Quantidade de tópicos em cada intervalo de pertinência tópica por estrato segundo STM.

Podemos adicionalmente observar o efeito da correlação de documentos e metadados na Figura 4.5, *Analyzing* na qual percebe-se um aumento das pertinências altas e médias em todos os estratos, mas não o suficiente para facilitar a identificação precisa dos documentos em cada estrato, o que nos leva a crer que esta dificuldade só poderá ser superada, no momento em que tivermos mais publicações classificadas manualmente nos estratos menores.

4.3 Classificação Bayesiana

Nesta seção usaremos a pertinência tópica para fazer a classificação em atenção a forma como YAU et al. (2014) faz em seu artigo, pegando o cálculo que fizemos da Pertinência tópica $P(e \mid t_j)$ para calcular a classificação.

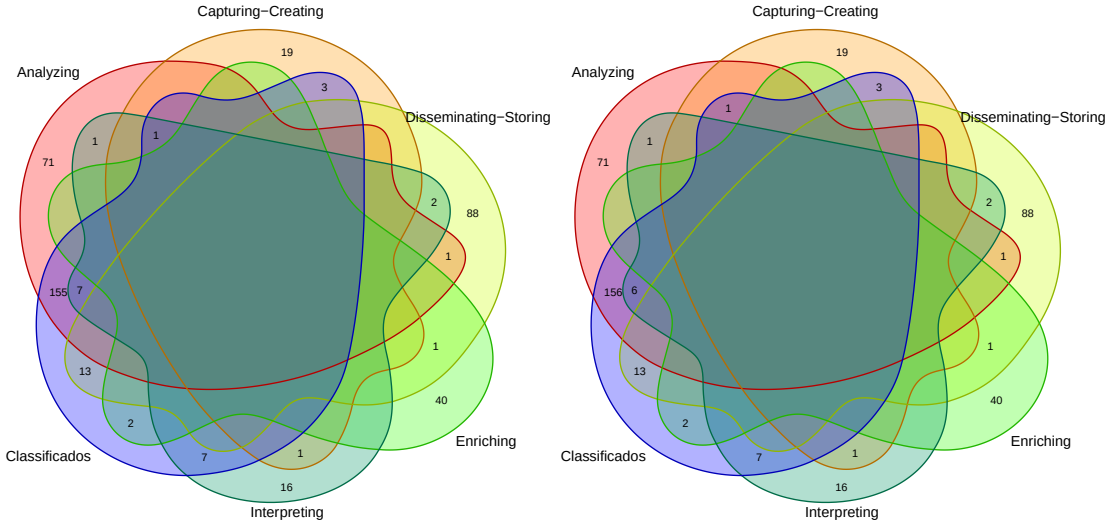
O nosso cálculo é muito parecido com o dos autores, sendo que a do artigo é um caso particular desse nosso caso aqui. Isso porque eles consideram cada tópico como inteiramente pertinente em apenas uma das classes, no nosso caso, em algum dos estratos da *TaDiRAH*. Por outro lado, nós usamos os percentuais de pertinência tópica equivalentes

à importância do tópico em cada estrato. Para que o nosso seja semelhante ao do artigo, basta que considerássemos pertinência de 100% em apenas um estrato, e 0% aos demais (YAU et al., 2014). Fazemos isso com o intuito de considerar a precisão na hora do cálculo da associação de cada estrato em cada publicação.

O funcionamento do método pode ser visto da seguinta forma. Digamos, hipoteticamente, que 50% de uma publicação é afetada pelo tópico 1. E que a pertinência tópica do tópico 1 seja de 90% para o estrato 1. Então é feita a multiplicação de 90% por 50% e teremos a fração da publicação que deve pertencer ao estrato 1. Além disso, a fração da publicação para o estrato 1 também sofre uma outra influência obtida pela soma da influência dos outros tópicos na participação desta publicação no estrato 1. O extrato em que essa soma for maior é o que a publicação será associada. Essa conta é explicitada abaixo na próxima subseção.

4.3.1 Método

Nesta seção explicamos o modelo bayesiano que aplicamos para classificar os textos segundo a *TaDiRAH* e a forma com que se avalia os resultados. Antes, porém, precisamos clarificar as definições que usaremos para que o nosso discurso flua mais naturalmente.



(a) Sobreposições classificadas incluídas múltiplas vezes.

(b) Publicações classificadas sem sobreposições.

Figura 4.6: Precisão e recolhimento de publicações classificadas $\mathbf{C}_{Analyzing}$.

Definições

1. $\mathbf{C}_e$: Quantidade de publicações que o nosso método classificou no estrato e.
2. $\mathbf{N}_e$: Quantidade de estratos que o nosso método classificou no estrato e, e ele realmente pertence ao estrato e.
3. $\mathbf{D}_e$: Quantidade de publicações que realmente estão no estrato e.
4. $\mathbf{D}'_\mathbf{e}$: Quantidade de publicações que realmente estão no estrato e sem as sobreposições.

5. $\mathbf{d}_i$: Cada publicação associada ao estrato.
6. $\mathbf{E_D}(\mathbf{d}_i)$: Conjunto de estratos da *TaDiRAH* que o documento $\mathbf{d}_i$ pertence.
7. $\mathbf{E_C}(\mathbf{d}_i)$: O estrato atribuído ao documento $\mathbf{d}_i$ pelo método bayesiano.
8. *Pontuação F*: $\mathbf{F}_e$, obtida como combinação de duas medidas: *precisão* e *recolhimento*.
9. *Precisão*: Fração classificada como e que acertou as publicações do estrato e, $\frac{N_e}{C_e}$. A precisão de *Analyzing* no exemplo da Figura 4.6 vale 0.8670213.
10. *Recolhimento*: Fração do alvo que foi acertada, $\frac{N_e}{D_e}$. O recolhimento de *Analyzing* no exemplo da Figura 4.6 vale 0.690678.

Construção do modelo: O modelo é construído durante a etapa de treino e teste, quando também fazemos a análise por tópicos durante a qual obtemos o termo $P(t_j \mid \mathbf{d}_i)$ e das pertinências tópicas $P(e \mid t_j)$. Ambas entrarão no cálculo da participação da publicação no estrato onde são feitas as distribuições $P(e \mid \mathbf{d}_i)$ associadas às proporções de observação, em cada publicação, de cada estrato.

Nesse sentido, temos como resultado, cada publicação $\mathbf{d}_i$ associada ao estrato resultante de

$$\arg\max_{e \in \mathbf{E}} \{P(e \mid \mathbf{d}_i)\} = \arg\max_{e \in \mathbf{E}} \left\{ \sum_{t_j \in \mathbf{T}} P(e \mid t_j) P(t_j \mid \mathbf{d}_i) \right\} \tag{4.1}$$

O resultado da aplicação de (4.1), é a obtenção de uma classificação descrita por $\{\mathbf{C}_e \subseteq \mathbf{D} \mid e \in \mathbf{E}$, sendo $|\mathbf{C}_e| = C_e$. A Figura 4.6 ilustra $\mathbf{C}_{Analyzing}$, com $C_{Analyzing} = 188$.

Avaliação

A Avaliação depende da *Pontuação F* onde, para cada estrato é calculada levando-se em conta quantidades de publicações distintas que explicaremos adiante. Tentaremos esclarecer ponto a ponto da formação da *Pontuação F* levando-se em conta dois mundos com pontos de vistas distintos, na tentativa de favorecer de um lado entendimentos mais voltados para a área computacional e por outro lado, os mais voltados para as áreas sociais.

A avaliação da classificação obtida consiste em verificar a semelhança entre $\mathbf{C}_e$ e $\mathbf{D}'_e$, para cada estrato e. Ou seja, verificamos a quantidade de publicações que o nosso método classificou segundo os estratos da *TaDiRAH* e comparamos com a quantidade de publicações que realmente estão em cada estrato sem as sobreposições.

Precisamos separar as publicações presentes no estrato e de suas sobreposições (repetições dessas publicações em outros estratos) para podermos fazer essas comparações. Dessa forma $\mathbf{D}'_e$ é definido como um subconjunto de $\mathbf{D}_e$ de forma que as sobreposições que aparecem na Figura 4.6a são desconsideradas nas publicações classificadas na Figura 4.6b.

Por isso, vamos detalhar melhor essas publicações sem as sobreposições, para entender bem, precisamos descrever mais precisamente a formação dessas publicações $\mathbf{D}'_e$. Sendo assim, definimos os dois tipos possíveis de associação entre cada publicação $\mathbf{d}_i$:

1. Primeiro tipo (publicações já classificadas do grupo zotero): O conjunto de estratos ao qual $\mathbf{d}_i$ pertence segundo a *TaDiRAH*, que denominamos $\mathbf{E_D}(\mathbf{d}_i) = \{e \in \mathbf{E} \mid \mathbf{d}_i \in \mathbf{D}_e\}$.
2. Segundo tipo (publicações classificadas pelo nosso método): O *estrato atribuído* a $\mathbf{d}_i$ *pelo método bayesiano*, para o qual usamos a notação $\mathbf{E_C}(\mathbf{d}_i)$.

A ideia em que se baseia a definição de $\mathbf{D}'_e$:

1. Hipótese 1 (Método acerta - conta a publicação como um acerto para o estrato da *TaDiRAH*): São mantidos as publicações de $\mathbf{D}_e$ cuja classificação determinada pelo método bayesiano vale e. O método bayesiano acertou a classificação de $\mathbf{d}_i$, em cujo caso essa publicação conta, para fins de avaliação, apenas para o estrato e (ou seja, $\mathbf{d}_i$ não é considerada como erro para os estratos em $\mathbf{E_D}(\mathbf{d}_i)$ diferentes de e).
2. Hipótese 2 (Método erra - Conta a publicação como um erro para todos os estratos da *TaDiRAH*): É um estrato que não pertence a $\mathbf{E_D}(\mathbf{d}_i)$. Corresponde ao caso em que o método bayesiano erra a classificação de $\mathbf{d}_i$, e por isso essa publicação é contada como erro em todos os estratos de $\mathbf{E_D}(\mathbf{d}_i)$.

A definição formal é $\mathbf{D}'_e = \{\mathbf{d}_i \in \mathbf{D}_e \mid \mathbf{E_C}(\mathbf{d}_i) = e \text{ ou } \mathbf{E_C}(\mathbf{d}_i) \notin \mathbf{E_D}(\mathbf{d}_i)\}$.

$D_e = |\mathbf{D}'_e|$. No exemplo da Figura 4.6, $D_{Analyzing} = 236$. Ou seja, a quantidade de publicações que realmente estão no estrato e com e sem as sobreposições são as mesmas.

$N_e = |\mathbf{C}_e \cap \mathbf{D}'_e|$. No exemplo da Figura 4.6, $N_{Analyzing} = 163$. Ou seja, a quantidade de publicações que o nosso método classificou no estrato e está integrado com a quantidade de publicações que realmente estão no estrato e sem as sobreposições.

A pontuação F de e resulta da combinação de *precisão* e *recolhimento* por média harmônica,

$$F_e = \frac{2}{\frac{C_e}{N_e} + \frac{D_e}{N_e}} = \frac{2N_e}{C_e + D_e}$$

A pontuação F de *Analyzing* no exemplo da Figura 4.6 vale 0.7688679.

Nas subseções seguintes, mostramos como aplicamos o método bayesiano usando os métodos LDA e STM.

4.3.2 Prevalência Tópica via LDA

Para o experimento e obtenção dos resultados mostrados na Tabela 4.1, aplicamos o método bayesiano (4.1) tendo as prevalências tópicas determinadas por meio da análise por tópicos com LDA. Separamos aleatoriamente as publicações em $\mathbf{D}$, definindo o treino e teste em 75% e 25% respectivamente.

Após a escolha aleatória das publicações, acrescentamos os textos descritivos para formar o corpus para a determinação das prevalências tópicas através de análise por tópicos via LDA e das pertinências tópicas. Em seguida, as publicações restantes são então utilizadas para a classificação segundo (4.1). Esse processo é repetido 10 vezes, e ao final é calculada a média dos resultados de todas as avaliações de cada repetição mostrada na Tabela 4.1.

Tabela 4.1: Atribuição das publicações aos estratos com o método eqrefeq:bayes e LDA (média sobre 10 amostragens).

Estrato	Precisão	Recolhimento	Pontuação
Analyzing	0.6836507	0.9679389	0.8004975
Capturing-Creating	0.2500000	0.0783333	0.1185714
Disseminating-Storing	0.8229229	0.6718835	0.7331295
Enriching	0.1500000	0.0253846	0.0433333
Interpreting	0.0000000	0.0000000	0.0000000

A *Pontuação F* nos dá o resultado da análise, que já explicamos na subseção anterior, como sendo a média harmônica da *Precisão* com o *Recolhimento* feitas para os 10 experimentos. Ou seja, em cada um dos experimentos são calculadas A *Precisão*, o *Recolhimento*, depois é feito a média harmônica para calcular a *Pontuação F*; depois de feitas todas as 10 medidas dos 10 experimentos diferentes é feita a média de cada uma dessas 3 grandezas e exposta aqui na Tabela 4.1.

Observando os valores da *Precisão*, do *Recolhimento* e da *Pontação F*, podemos notar claramente que os estratos *Analyzing* e *Disseminating-Storing* possuem resultados satisfatórios em detrimento dos demais estratos. O que significa dizer que os resultados mais precisos aparecem nos estratos mais fortemente representados por tópicos como já vimos lá na seção da *Pertinência Tópica*. Esses resultados endossam nossa afirmação de que essa alteração se dá em consequência do desbalanceamento da amostragem.

4.3.3 Prevalência Tópica via STM

Aplicando o STM de acordo com (4.1) obtemos resultados um pouco melhores, devido às correlações e os agrupamentos, porém ainda não foram satisfatórios. Mais uma vez *Analyzing* e *Disseminating-Storing* tem bom desempenho, porém os demais estratos ficam insatisfatórios. Isso devido mais uma vez a pertinência tópica que é alta nos estratos com maior quantidade de publicações e baixa nos estratos com poucas publicações. Os resultados estão demonstrados na Tabela 4.2.

Sendo assim, o desbalanceamento é um fator extremamente importante que afeta grandemente os resultados, fazendo com que os estratos maiores tenham mais força para classificar as suas publicações e ainda carregar junto às publicações dos estratos menores. Os tópicos perdem força quando não tem representatividade suficiente dada pela quantidade de publicações classificadas artesanalmente.

4.4 Classificação LDA Supervisionado

Esta seção mostra o uso do classificador LDA Supervisionado (BLEI; MCAULIFFE, 2010) a fim de compararmos com o método bayesiano proposto, como mostrado no Capítulo 3, onde também mostramos que para usar o LDA Supervisionado (sLDA) precisamos tomar algumas providências para contornar a questão das sobreposições porque o sLDA pressupõe que não haja sobreposição de estratos. O que acontece por causa do método

Tabela 4.2: Atribuição das publicações aos estratos com o método eqrefeq:bayes e STM (média sobre 10 amostragens).

Estrato	Precisão	Recolhimento	Pontuação
Analyzing	0.7280099	0.9325254	0.8157659
Capturing-Creating	0.6500000	0.1920238	0.2841270
Disseminating-Storing	0.7882040	0.8176463	0.7967825
Enriching	0.4833333	0.1544406	0.2228480
Interpreting	0.2000000	0.0291667	0.0507937

utilizado para gerar o dado de atribuição de tópicos, que é construído enquanto ele vai jogando os dados e construindo os documentos.

A forma que ele constrói esse dado depende de quantas vezes cada tópico caiu durante a construção do documento e a forma que ele combina isso é através de uma Regressão Logística. Uma regressão, por hipótese, pressupõe que não há sobreposições entre as classes. Então esse método supervisionado não se adequa aos casos com sobreposição de estratos.

Classificação Binária

A fim de lidar com as sobreposições, a abordagem adotada consiste em realizar diversas rodadas de classificação, cada qual com um modelo sLDA próprio a um estrato. Cada rodada é dedicada ao estrato próprio do modelo utilizado e o objetivo é decidir, para cada publicação a classificar, se ela pertence ou não ao estrato considerado. Desta maneira, cada publicação pode ser classificada em diversos estratos, fato que corresponde ao perfil da *TaDiRAH*:

1. O modelo sLDA próprio de um estrato é construído tomando todas as publicações desse estrato como uma classe e as publicações *exclusivas* dos demais estratos como a segunda classe:

- Pegamos todas as publicações que estão em um determinado extrato;
- Retiramos essas publicações dos outros extratos quando houver sobreposição;
- Fazemos uma classificação (por estrato) dessas publicações que estão neste estrato contra as publicações que estão nos outros estratos;
  - A classificação sempre será binária;
  - Neste tipo de classificação a cada rodada decidimos para cada publicação testada, se ela está neste estrato, ou não;
- Repetimos a operação com outro estrato, da mesma forma, até que tenhamos feito com todos os estratos e todas as publicações.

Exemplificando com o teste de *Analyzing* onde queremos saber se a publicação está no estrato *Analyzing* ou não está? Se sim, classificamos em *Analysing*, senão, não será classificado nesta rodada. Então passamos para *Disseminating-Storing*, até completar todos os 5 estratos da nossa análise.

As publicações próprias de um estrato e são todas as publicações em $\mathbf{D}_e$, enquanto as an-

Tabela 4.3: Publicações próprias aos estratos da *TaDiRAH* e exclusivamente em estratos distintos.

	Própria	**Antípoda**
Analyzing	272	290
Enriching	61	501
Capturing-Creating	59	503
Disseminating-Storing	145	417
Interpreting	57	505

Tabela 4.4: Atribuição das publicações aos estratos com o método LDA supervisionado (média sobre 10 amostragens).

Estrato	**Precisão**	**Recolhimento**	**Pontuação**
Analyzing	0.6753337	0.3712988	0.4703681
Capturing-Creating	0.2081424	0.2382143	0.1727796
Disseminating-Storing	0.6470353	0.6520086	0.6039828
Enriching	0.3360323	0.3542541	0.2781453
Interpreting	0.1238724	0.3202273	0.1449815

típodas são as publicações exclusivas dos demais estratos. As quantidades de publicações próprias e antípodas dos estratos são apresentadas na Tabela 4.3.

Por definição, os conjuntos de publicações próprias e antípodas são disjuntos. Ou seja, não há sobreposição de estratos. Dessa forma, o LDA supervisionado pode ser utilizado sem que haja comprometimento da regressão logística envolvida. Além disso, foram acrescentados os textos descritivos dos estratos da *TaDiRAH* junto com as publicações dos estratos no intuito de reforçar os temas e melhorar a classificação.

Resultados

A Tabela 4.4 demonstra os resultados da classificação binária usando o LDA Supervisionado. Se compararmos os resultados com os resultados do método bayesiano da seção anterior, podemos notar que o desempenho da classificação binária ficou bem inferior. Isto se dá por causa do desbalanceamento.

Ao fazermos a manobra de classificar um estrato contra todos os outros aumentamos ainda mais o desbalanceamento porque classificamos um contra o somatório dos demais. Para verificar esta afirmação basta olharmos na tabela onde em cada estrato identificado na coluna *Própria* temos as publicações próprias dos estratos, enquanto que na coluna *Antípoda* temos o somatório de todos os outros estratos, sendo 5 estratos no total, teremos sempre 1 estrato contra 4 outros. Essa diferença só não é demasiada em *Analyzing*, nos outros essa diferença é de 3 a 9 vezes.

A conclusão geral a que chegamos é que o método Bayesiano é bom, porém está sendo atrapalhado pelo desbalanceamento, e que não dá pra resolver o problema da classificação de publicações usando a *TaDiRAH* direito sem tratar o desbalanceamento. Com um

método para resolver ou amenizar o desbalanceamento podemos conseguir melhorar estes resultados. Porém o ideal é realmente ter mais documentos classificados de todos os estratos para ficar mais nítida e equilibrada a nossa classificação.

5 TRATAMENTO DE DESBALANCEAMENTO DE ESTRATOS

Neste capítulo descreveremos a nossa estratégia para lidar com o problema do desbalanceamento de publicações por estrato que abordamos no Capítulo 4. A estratégia que apresentamos é baseada na geração de amostras sintéticas dos estratos minoritários e os experimentos que referenciam a abordagem.

Após a geração de publicações sintéticas textuais, que é aplicada ao contexto do método proposto no Capítulo @ref (sec:assist), apresentamos os resultados da classificação Bayesiana com as publicações sintéticas aplicadas ao LDA (BLEI; NG; JORDAN, 2003) e no STM (ROBERTS et al., 2013).

Por fim, mostraremos como efetuamos o treinamento do classificador sLDA (BLEI; MCAULIFFE, 2010), no contexto das sobreposições seguindo o passo-a-passo observado no Capítulo 3 para a fase de treinamento; a exigência de publicações já classificadas e a influência do desbalanceamento na quantidade de amostras por estrato sobre a classificação.

5.1 Abordagens

Existe uma vastíssima literatura que mostra como podemos lidar com amostras desbalanceadas em problemas de classificação, como mostra bem a revisão bibliográfica feita em EBENUWA (2019). Porém os tratamentos de desbalanceamento em problemas de classificação normalmente lidam com problemas de atributos numéricos ou para amostras vetoriais de múltiplos atributos: variações de SVM, árvore de decisão, redes neurais e seleção de atributos.

Nas abordagens vetoriais cada amostra pode ser vista como um ponto em um certo espaço dimensional em que cada dimensão desse espaço é um dos atributos, em geral são atributos numéricos. Onde os pontos referentes às amostras formam uma estrutura em um espaço numérico, que será analisada para, a partir daí, tentar extrair padrões para serem usados com os métodos.

Porém, essas abordagens não são adaptáveis a métodos de análise por tópicos e, também não está dentro do escopo desta dissertação uma possível adaptação de algum método de resolução de problemas de amostras vetoriais de múltiplos atributos, tais como:

1. Aprendizagem Sensível ao Custo (*Cost-Sensitive Learning – CSL*) (LING; SHENG, 2010; SUN et al., 2007): Onde, com base em uma metodologia, se avalia e atribui pesos diferentes às classes segundo as suas sensibilidades a erros de classificação estabelecidas a partir de certas propriedades. Nesta abordagem a precisão final não é mais importante do que o custo de prever erroneamente um falso negativo (ex.: um falso negativo de um exame de um paciente com câncer, é considerado pior do que um falso positivo de câncer em um exame de um paciente saudável). No contexto deste trabalho, o nosso conjunto de publicações da *TaDiRAH* extraídos de DARIAH-UE CONSORTIUM (2012), quanto menor for a sua quantidade de amostras, mais sensível ao erro seria o estrato, sendo melhor ter uma acurácia um pouco mais baixa, porém detectando o estrato minoritário, do que ter uma acurácia

alta mascarada pela grande quantidade de amostras do estrato majoritário.

2. Vizinhança mais próxima (*K-Nearest Neighbors – KNN*) (ALTMAN, 1992): Onde uma quantidade de publicações fixadas em um espaço numérico servem de parâmetro de distanciamento para classificar uma outra publicação **d** como mais próxima, segundo alguma métrica de similaridade. No nosso contexto é inviável, pois teríamos que converter os textos em dados numéricos, aplicar o método e retornar para o texto. Abordaremos um pouco mais sobre os perigos dessa ida e volta ao falarmos das amostras sintéticas textuais.

3. Coletivo de algoritmos (*Ensemble Classification*) (BRAMER, 2013): Onde são empregados sucessivos algoritmos sincronizados aprendendo em conjunto para funcionar em cooperação combinando suas previsões de alguma forma, por exemplo, no contexto do nosso trabalho, poderia ser feitas análises em camadas de saída e entrada, onde os dados de saída do algoritmo anterior servem como dados de entrada para o próximo algoritmo da fila de execução. A idéia dessa cooperação é aprimorar as capacidades preditivas individuais dos algoritmos. Espera-se uma espécie de sinergia, onde o trabalho do conjunto dos algoritmos tenham um desempenho melhor do que qualquer um deles trabalhando individualmente.

Na seção seguinte, mostramos a abordagem que adotamos com o objetivo de diminuir o desbalanceamento dos estratos a partir da geração de textos sintéticos, onde usamos a técnica conhecida como *SMOTE*, sigla em inglês para *Técnica de Reforço por Amostras Sintéticas* (CHAWLA et al., 2002), uma técnica normalmente usada em problemas envolvendo atributos numéricos.

A estratégia de geração de publicações sintéticas que adaptamos é baseada na idéia descrita em JAPKOWICZ (2000) onde vemos que a simples repetição das amostras originais (autênticas) para formar novos textos complementares de reforço dos estratos minoritários não surte o efeito necessário de aumentar o reconhecimento desses estratos pelo algoritmo de classificação, sendo assim, usaremos as características espaciais de decisão que definem o perfil dos estratos minoritários na geração dos textos sintéticos destes estratos.

5.2 Amostras Sintéticas Textuais

Há duas abordagens, não exclusivas, que são empregadas para balancear as quantidades de amostras entre classes desbalanceadas, a saber:

1. *Subamostragem*: Na qual reduz-se a quantidade de amostras das classes majoritárias ou;

2. *Superamostragem*: Que corresponde a aumentar a quantidade da(s) classe(s) minoritária(s).

A subamostragem depende essencialmente do critério de escolha das amostras a guardar na classificação. Esse critério pode ser tanto uma escolha aleatória (CHAWLA et al., 2002) como escolhas dirigidas por análises da distribuição de classes, tal como o método de *Subamostragem Radial (RBU)* (KOZIARSKI, 2020), adaptado da *Superamostragem Radial (RBO)* (KOZIARSKI; KRAWCZYK; WOŹNIAK, 2017), assim como estratégias de agrupamentos de amostras mais significativas (SONG et al., 2016).

Ambas as abordagens não podem ser feitas sem os devidos cuidados porque senão podemos

aumentar ainda mais o problema. Ao diminuirmos a quantidade da classe majoritária sem os devidos cuidados, podemos prejudicar a sua classificação, sem necessariamente melhorar a classificação da classe majoritária. Uma vez que não temos muitas publicações classificadas por especialistas da área em alguns dos estratos, poderíamos ficar com pouca informação (poucas amostras de treino) para treinar o algoritmo, o que implicaria em um treino ruim, e algoritmo mal treinado, classifica mal.

Se, por outro lado, o aumento da quantidade de publicações da classe minoritária por replicação supervaloriza as características dessas publicações, provocando o efeito de sobreajuste (denominado *overfitting* na literatura em língua inglesa) por tornar a classe muito específica, o que faz o algoritmo ter a tendência de classificar no estrato minoritário somente as publicações idênticas, ou muito semelhantes, às que foram replicadas.

Uma forma que despertou o nosso interesse é o aumento da quantidade de amostras através de técnicas de superamostragem, gerando publicações sintéticas. O objetivo é fazer com que as publicações sintéticas guardem as características das publicações autênticas sem com isso ser uma réplica delas. O benefício de usar publicações sintéticas é que o algoritmo pode aprender melhor quais as características das publicações de um estrato específico, que o difere dos demais, sem decorar aquele estrato como constituído exclusivamente pelas amostras de treino. Porém, os melhores métodos existentes para gerar amostras sintéticas que encontramos na literatura são voltadas para amostras numéricas.

Dos métodos para gerar textos sintéticos, encontramos duas abordagens que poderíamos usar no nosso experimento:

1) A primeira abordagem é transformar o texto para vetores, aplicar a técnica de replicar vetores, e depois voltar novamente os vetores para textos.

Problema: Como manter as propriedades da relação desses vetores com os textos, e vice e versa? Isso porque os vetores gerados pelo método podem fazer sentido quando estiverem como vetores, por estarem no mesmo envelope convexo, mesma região, no mesmo perfil, possui as similaridades estruturais, e reforça o perfil deles entre os vetores, mas podem perder o sentido quando transformados em textos novamente;

2) A segunda abordagem é gerar textos aleatoriamente.

Problema: Como guardar padrões das publicações autênticas da classe para reforçar esses padrões nos textos sintéticos? Os textos da classe minoritária são poucos, por isso o padrão desses textos é difícil de ser percebido, precisamos capturar o perfil do texto para que possamos gerar textos com o mesmo perfil dos existentes e assim reforçar esse padrão na classe de origem.

Os classificadores bayesianos e sLDA tratam cada publicação como uma cesta de palavras. Por essa razão, não é necessário gerar publicações propriamente ditas, e sim gerar uma cesta de palavras que preserve o padrão da cesta de palavras do estrato correspondente. Ou seja, a ordem das palavras nesta cesta não é importante, o que importa mesmo é determinar as palavras candidatas a compor a publicação sintética.

Neste contexto, a questão fundamental é a seguinte: Como efetuar a escolha aleatória das palavras para compor a cesta de uma publicação sintética, e qual a quantidade de cada uma dessas palavras?

Naturalmente, a resposta a essa questão deve ser de tal forma que uma publicação sintética

gerada mantenha o perfil das publicações daquele estrato. Então, manter o padrão é descobrir quais as palavras e a forma de fazer o sorteio para que as publicações sintéticas tenham o mesmo perfil das autênticas.

Nossa Proposta Tendo em vista que o método LDA de análise por tópicos desvenda as variáveis latentes de um modelo generativo, propomos usar este mesmo modelo para construir as publicações sintéticas. Mais precisamente a proposta consiste em:

1. Aplicar o LDA para construir o modelo generativo de cada estrato e;
2. usá-lo para gerar publicações com perfis similares aos das publicações que originaram o modelo.

O fundamento para esta proposta está no fato de que se soubermos os padrões das palavras dentro das publicações autênticas, poderemos gerar novas publicações com tópicos semelhantes a partir de sorteios aleatórios usando os padrões descobertos e as palavras que compõem as publicações autênticas.

Sendo assim, passamos agora a descrever a abordagem proposta em duas etapas:

Primeira Etapa: 1. Pegamos as publicações autênticas, e dividimos nos grupos referentes aos estratos, quer dizer, separamos os estratos, de forma que agora no lugar de termos um único corpus com todas as publicações autênticas de todos os estratos, passamos a ter um corpus para cada estrato;

2. Fazemos uma análise por tópicos prévia de cada estrato individualmente usando o LDA, construindo um modelo generativo por estrato. Com isso estabelecemos o perfil de cada estrato separadamente. Ao fazer essa análise remontamos o modelo LDA, já que o objetivo do método LDA é montar o modelo que supostamente gerou, ou pelo menos, se aproxima do modelo que gerou aqueles textos. Ou seja, remontamos os modelos com os perfis das publicações autênticas de cada estrato.

Remontar o modelo significa, estabelecer 3 parâmetros:

a. Parâmetro $\boldsymbol{\theta}$: As relevâncias tópicas;
b. Parâmetro $\boldsymbol{\beta}$: As relevâncias semânticas e;
c. Parâmetro $\boldsymbol{\alpha}$: O parâmetro da distribuição de Dirichlet, que é a distribuição que gera os dados de tópicos, que gera as relevâncias tópicas.

Esses parâmetros são determinados para cada um dos estratos nesta primeira etapa.

Segunda Etapa:

3. Usamos o modelo generativo de cada estrato representado pelos parâmetros de distribuição $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$ para gerar as publicações sintéticas desse estrato baseado na premissa de que em uma análise LDA os textos são gerados a partir de um modelo gerador, sendo que o método reconstrói esse modelo gerador. Logo, agora que temos o modelo gerador, podemos gerar novos textos com ele.

Na seção seguinte, mostraremos detalhadamente (a) *A Primeira Etapa*: Geração do modelo, e principalmente como é feita e; (b) *A Segunda Etapa*: Geração de textos propriamente dita, após a determinação do modelo.

Nas seções seguintes mostraremos como, nos modelos de classificação bayesiano e sLDA, incluímos as publicações sintéticas obtidas para aumentar a quantidade de publicações

das classes minoritárias.

5.3 Experimentos de Geração de Publicações Sintéticas

Na seção anterior deixamos claro que precisamos determinar os parâmetros $\boldsymbol{\alpha}$ e $\boldsymbol{\theta}$ para que o modelo descrito na Seção 3.1.3, onde descrevemos o modelo do LDA, possa ser efetivamente usado para a geração de publicações sintéticas.

Nesta seção alinharemos os detalhes da determinação desses dois parâmetros pela aplicação do método LDA e a subsequente reconstrução desses parâmetros por inferência no resultado da aplicação do LDA.

Parâmetro $\boldsymbol{\alpha}$: Regula a distribuição conjunta de Dirichlet na determinação do dado de tópicos utilizado em cada publicação sintética.

Parâmetro $\boldsymbol{\theta}$: O próprio dado de tópicos, representado por um vetor tendo como elementos as probabilidades de ocorrência de cada tópico.

Sendo assim temos como relação de dependência: Um parâmetro $\boldsymbol{\alpha}$ para cada estrato, que gera um parâmetro $\boldsymbol{\theta}$ para esse mesmo estrato, que por sua vez gera cada publicação sintética, ou cesta de elementos léxicos exclusivo para cada estrato.

De acordo com a quantidade das publicações autênticas, escolhemos aleatoriamente a quantidade de amostras sintéticas por estrato segundo uma distribuição de Poisson (BLEI; NG; JORDAN, 2003; MANNING; SCHÜTZE, 1999).

Os perfis das relevâncias tópicas autênticas $\boldsymbol{\theta}_a$ e das relevâncias tópicas sintéticas $\boldsymbol{\theta}_s$ são aderentes, ou integradas entre si em razão das relevâncias tópicas de reforço.

Relevâncias Tópicas de Reforço São as repetições das relevâncias tópicas autênticas replicadas na formação das publicações sintéticas com o objetivo de reforçar o perfil das publicações autênticas nas publicações sintéticas.

Relevâncias Tópicas de Diversificação São relevâncias adicionais, usadas exclusivamente em uma publicação sintética e que não são advindas de repetição das publicações autênticas. Ou seja, são exclusivas e usadas com o objetivo de diversificar as publicações sintéticas.

A diversificação das relevâncias tópicas sintéticas é determinada para atuar como pequenas nuances para que não haja a simples repetição de padrões que provocam o efeito de *overfitting*, já explicado na Seção 5.2, conforme Figura 5.1 e Tabela 5.1.

Observamos também que as relevâncias tópicas sintéticas $\boldsymbol{\theta}_s$ tendem a fortalecer a relevância tópica dos tópicos mais relevantes por transferência direta provocada pela repetição nas relevâncias tópicas de reforço, e diversificando com as relevâncias tópicas de diversificação nos tópicos menos relevantes.

Porém, apesar do reforço das relevâncias tópicas mais relevantes em $\boldsymbol{\theta}_a$, dado o desvio padrão, uma menor correlação entre as médias de $\boldsymbol{\theta}_a$ e $\boldsymbol{\theta}_s$ pode acontecer indicando a oscilação dessas relevâncias tópicas aumentando ou diminuindo uma em relação a outra.

O Algoritmo SINTÉTICAS$(e, D_a, \boldsymbol{\theta}_a, \boldsymbol{\alpha}_a, \mathbf{M}_a, \delta, D_s, \mathbf{D}_a, \boldsymbol{\beta}_z)$ mostra como gerar as publicações sintéticas de um determinado estrato, dado o seu modelo. O algoritmo será aplicado

individualmente em cada estrato minoritário, com o objetivo de estabelecer o seu perfil, utilizando os seguintes parâmetros:

1. *Identificação do Estrato* e: No caso, o estrato minoritário que está sendo trabalhado;
2. *Quantidade de publicações autênticas* D_a: Referente ao estrato e identificado acima. Esse parâmetro é utilizado para estabelecer o conjunto das possíveis relevâncias tópicas das amostras sintéticas, aquelas que serão repetidas para fortalecer o padrão do perfil das publicações autênticas;
3. *Corpus das publicações autênticas* $\mathbf{D}_a$ do estrato e;
4. *Relevâncias tópicas das publicações autênticas* $\boldsymbol{\theta}_a$: obtidas com a análise por tópicos prévia do corpus $\mathbf{D}_a$ das publicações autênticas do estrato e;
5. *Análise por tópico prévia*: Consideradas somente os elementos léxicos mais relevantes que ocorrem pelo menos em 5 publicações e ao menos 10 vezes ao todo, no corpus das publicações autênticas $\mathbf{D}_a$ do estrato e. A análise é feita tendo $\min\{30, 0.9D_a\}$ como quantidade de tópicos da análise das publicações autênticas buscando $\boldsymbol{\alpha}$_a que privilegie poucos tópicos.

6. *Parâmetro da distribuição de Dirichlet* $\boldsymbol{\alpha}_a$: Obtido na análise por tópicos prévia do corpus $\mathbf{D}_a$;

7. *Fator de diversificação* $\delta \geq 1$: Indica a fração adicional de relevâncias tópicas sintéticas que são obtidas a partir do parâmetro da distribuição de Dirichlet $\boldsymbol{\alpha}_a$;
8. *Fator de Diversificação* δ: O percentual de diversificação das relevâncias tópicas sintéticas dada a partir das relevâncias tópicas autênticas. Ou seja, a quantidade de relevâncias tópicas de diversificação que será usada na construção das publicações sintéticas;
9. *Quantidade de publicações sintéticas* $\boldsymbol{\theta}_s$: Devem ser geradas, de forma a que o tamanho do corpus das publicações sintéticas sejam dadas levando-se em conta o fator de diversificação atuante no corpus das publicações autênticas $D_s = \delta D_a$;
10. *Relevâncias tópicas de reforço* $\boldsymbol{\theta}_s[1, \ldots, D_a]$: Usadas na construção das publicações sintéticas com o objetivo de reforçar o perfil das publicações autênticas. Seus valores são idênticos às correspondentes nas relevâncias tópicas autênticas $_a$;
11. *Relevâncias tópicas de diversificação* $\boldsymbol{\theta}_s[D_a + 1, \ldots, D_s]$: Usadas para gerar publicações sintéticas com perfil de combinação de tópicos diverso em relação às publicações autênticas. Para que elas sejam coerentes com os tópicos estabelecidos pelo modelo, elas são geradas do próprio Dirichlet usando o parâmetro $\boldsymbol{\alpha}_a$;
12. *Metadados* $\mathbf{M}_a$: São os tipos de metadados que usamos nas publicações sintéticas. Seus valores são definidos por $\langle$`ItemType`, `PublicationYear`, `Class`$\rangle$;

13. *Parâmetro da relevância semântica* $\boldsymbol{\beta}_z$;

14. *Corpus autêntico* $\mathbf{D}_a$.

O Algoritmo Sintéticas$(e, D_a, \boldsymbol{\theta}_a, \boldsymbol{\alpha}_a, \mathbf{M}_a, \delta, D_s, \mathbf{D}_a, \boldsymbol{\beta}_z)$ recebe alguns parâmetros de entrada. Basicamente temos como parâmetros:(1) o parâmetro da distribuição Dirichlet $\boldsymbol{\alpha}_a$, (2) relevâncias semânticas $\boldsymbol{\beta}_z$ e (3) relevâncias tópicas do estrato $\boldsymbol{\theta}_a$ que definem o modelo que vai gerar.

Algoritmo: SINTÉTICAS$(e, D_a, \boldsymbol{\theta}_a, \boldsymbol{\alpha}_a, \mathbf{M}_a, \delta, D_s, \mathbf{D}_a, \boldsymbol{\beta}_z)$: geração de publicações sintéticas de um estrato

para todo $d \in \{1, \dots, D_a\}$ **faça**
    Relevâncias tópicas de reforço $\boldsymbol{\theta}_{s,d} \leftarrow \boldsymbol{\theta}_{a,d}$
    $\xi_d \leftarrow |\mathbf{D}_{a,d}|$
para todo $d \in \{D_a + 1, \dots, \delta D_a\}$ **faça**
    Relevâncias tópicas de diversificação $\boldsymbol{\theta}_{s,d} \sim Dirichlet(\boldsymbol{\alpha}_a)$
    $\xi_d \leftarrow \overline{|\mathbf{D}_a|}$
Criar corpus vazio $\mathbf{D}_s \leftarrow \emptyset$
para todo $d \in \{1, \dots, D_s\}$ **faça**
    Criar a cesta de elementos léxicos vazia $\mathbf{d} \leftarrow \emptyset$
    Escolher dado de tópicos $t \sim Uniforme(1, \delta D_a)$
    Escolher quantidade de elementos léxicos $N \sim Poisson(\xi_t)$
    para $n \leftarrow 1, \dots, N$ **faça**
        Escolher tópico $z \sim Multinomial(\boldsymbol{\theta}_{s,t})$
        Escolher elemento léxico w usando a probabilidade multinomial β_z
        Acrescentar w à cesta $\mathbf{d} \leftarrow \mathbf{d} \cup \{w\}$
    se $t \leq D_a$ **então** ▷ Acrescentar $\mathbf{d}$ ao corpus com os metadados
        $\mathbf{D}_s \leftarrow \mathbf{D}_s \cup (\mathbf{d}, \mathbf{M}_{a,t})$
    senão
        $\mathbf{D}_s \leftarrow \mathbf{D}_s \cup (\mathbf{d}, \{\texttt{Sintético}, 2022, e\})$
retorna $\mathbf{D}_s$

Agora temos os parâmetros adicionais: (4) a identificação do estrato e, (5) a quantidade de publicações autênticas que esse estrato tem D_a; (6) quais são os metadados $\mathbf{M}_a$ (tipo de publicação, ano e estrato); (7) os parâmetros que controlam o que a gente vai gerar D_s e (8) o fator de diversificação δ.

Primeiramente pegamos o modelo original oriundo da primeira etapa descrita na parte 2 da *Primeira Etapa* da Seção anterior como análise por tópico prévia, onde obtivemos o modelo original, então, aqui no algoritmo esse modelo original sempre aparece com um subscrito $(_a)$ identificando-os como autênticos. Fazemos algo semelhante para identificar as sintéticas usando o subscrito $(_s)$

Com esse modelo em mãos montamos o modelo que vamos efetivamente usar, que é muito parecido com o modelo original, porém, a parte mais importante é que fazemos uma pequena mudança para gerar uma diversificação.

Como fazemos

1. As relevâncias semântica $\boldsymbol{\beta}_z$: Não tocamos, usamos exatamente como ela veio da análise por tópicos prévia;

2. O parâmetro de Dirichlet $\boldsymbol{\alpha}_a$: Também não tocamos, usamos exatamente como veio da análise por tópicos prévia;

3. As relevância tópicas $\boldsymbol{\theta}_a$: Aqui sim. Agimos aqui para gerar alguma diversificação com relação às publicações autênticas com o intuito de não gerar coisas muito repetidas com as publicações autênticas, então geramos alguma diversificação; i. Então

para cada publicação sintética vamos escolher uma relevância tópica para ela de um conjunto de opções, para isso temos um conjunto de dados de tópicos lá. ii. Esses dados de tópicos, que são as relevâncias tópicas que vieram da análise por tópico prévia. Simplesmente repetimos e chamamos de *relevância tópica de reforço* para gerar publicações sintéticas que reforçam aquele perfil que já aparece nas autênticas; iii. Depois vamos usar também algumas outras relevâncias tópicas adicionais que chamamos de *relevâncias tópicas de diversificação* para gerar alguma diversificação no perfil das publicações sintéticas. A quantidade de diversificação é definida no parâmetro δ, onde a quantidade total de relevâncias tópicas a utilizar é igual a δ vezes D_a. Logo, a quantidade de relevâncias tópicas de diversificação é igual a δ menos 1, vezes D_a. Obs.: Nos experimentos usamos $1,1$, ou seja 10% a mais (pegamos as relevância tópicas oriundas da análise prévia, e adicionamos mais 10%).

Passamos agora a explicar o funcionamento do algoritmo:

Nas linhas 1 e 2 do algoritmo: As primeiras D_a são exatamente aquelas da análise por tópico prévia das publicações autênticas enquanto que as seguintes, nós criamos novas, para a diversificação, do mesmo jeito que o modelo LDA determinou: usamos o Dirichlet com o parâmetro que o LDA determinou, então procedemos do mesmo jeito que o modelo LDA gera. Ficando ao todo com δ vezes D_a relevâncias tópicas (uma quantidade a mais do que D_a. Ao fazer isso, estabelecemos θ_s. Agora, por já possuirmos também o β, temos todos os dados que precisamos para gerar as publicações sintéticas.

Na linha 3 do algoritmo $\xi_d \leftarrow |\mathbf{D}_{a,d}|$: Montamos o conjunto de relevâncias tópicas que usaremos.

Na linha 7 do algoritmo: Criamos um corpus vazio para receber essas publicações sintéticas. O parâmetro D_s define a quantidade de publicações sintéticas que vamos gerar.

A partir da linha 8 do algoritmo:

[**Linha 8**] Vamos repetir os procedimentos D_s vezes;

[**Linha 9**] Pega uma cesta vazia;

[**Linha 10**] Escolha do dado de tópicos que usaremos: Para isso, em uma distribuição uniforme, escolhemos (ao acaso) uma das relevâncias tópicas e usamos: Pode ser uma relevância tópica de reforço ou de diversificação;

O dado de elementos léxicos é o β que recebemos da primeira etapa;

[**Linha 11**] Definimos o tamanho dessa publicação sintética: Faremos igual ao LDA que usa uma outra distribuição chamada Poisson, que tem um parâmetro ξ_t:

[**Linha 11**].*[1]* Então se a relevância tópica t que sorteamos na linha 10, for uma relevância tópica de reforço, que está associada a uma publicação autêntica, o parâmetro ξ_t é o tamanho dessa publicação autêntica associada a essa relevância tópica de reforço.

[**Linha 11**].*[2]* Mas, se por outro lado, for uma relevância tópica de diversificação, que não tem nenhuma publicação autêntica associada a ela, então usamos como parâmetro ξ_t a média dos tamanhos de todas as publicações autênticas desse estrato. Esses valores estão colocados nas linhas 2 e 3 do algoritmo, onde: *[Linha 2]* para as distribuições de reforço colocamos lá naquele parâmetro; *[Linha 3]* ξ_d, o tamanho do corpus autêntico correspondente aquela relevância tópica $|\mathbf{D}_{a,d}|$.

[**Linha 11**]*.[2].[A]* Então, se olharmos na *linha 2 do algoritmo*, notamos que a relevância tópica autêntica de índice D $\theta_{a,d}$ está associada a publicação autêntica de índice d $d_{a,d}$, então colocamos no parâmetro *[Linha 3]* ξ_d o tamanho dessa publicação autentica de indice d $|\mathbf{D}_{a,d}|$;

[**Linha 11**]*.[2].[B]* Já na linha 5 do algoritmo, para as relevâncias tópicas de diversificação, colocamos no parâmetro ξ_d da linha 6 a média dos tamanhos das publicações nos tópicos $\overline{|\mathbf{D}_a|}$.

Uma vez escolhido o dado de tópicos e o tamanho das publicações sintéticas. Repetimos passos das *linhas 13, 14 e 15 do algoritmo*, para o tamanho escolhido (até preencher a cesta de palavras):

[**Linha 13**] Jogamos o dado de tópicos, sai o tópico;

[**Linha 14**] Jogamos o dado de elementos léxicos daquele tópico, escolhe o elemento léxico e;

[**Linha 15**] Colocamos esse elemento léxico na cesta de palavras;

[**Linha 16**] Agregamos os metadados. Obs.: Os metadados serão usados somente no método STM, mas mesmo assim precisamos colocá-los.

[**Linha 16**]*[1]* Então se na *[Linha 10]* escolhemos uma relevância tópica t de reforço, entre os primeiros D_a, então usamos *[Linha 17]* os mesmos metadados da publicação autêntica associada a relevância tópica de reforço que foi usada $M_{a,t}$, que vem com o parâmetro descrito no cabeçalho do algoritmo;

[**Linha 16**]*[2]* Senão, se escolhemos uma relevância tópica de diversificação, *[Linha 18]* Criamos os metadados *[Linha 19]* específicos para essas publicações sintéticas de diversificação. Então, todas as publicações sintéticas desse estrato e que foram geradas com metadados de diversificação vão ter o mesmo valor ou conjunto de metadados, e vão cair no mesmo grupo do STM. Todas terão o mesmo conjunto de valores $\{\texttt{Sintético}, 2022, e\}$, ou seja: (a) TIPO: *sinttico*, (b) ANO: 2022 e (c) ESTRATO: O parâmetro e.

Sendo assim:

Após a análise de todas as publicações autênticas identificamos o comportamento médio da probabilidade de ocorrência de cada tópico comparativamente com as publicações sintéticas. O comparativo pode ser visualizado na Figura 5.1 que traça um panorama de como se comporta de forma geral a nossa geração de publicações sintéticas, apenas para que possamos visualizar como esse processo se comporta:

1. Lembrando que esse é um experimento de geração de publicação sintéticas a partir de todo o conjunto de publicações autênticas, onde para cada estrato, pegamos todas as publicações autênticas daquele estrato para realizar esse processo de geração de sintéticas;

2. Comparando as publicações autênticas com as publicações sintéticas, teremos o grupo de autênticas (simbolizadas pelas barras verticais mais próximas da cor laranja) e o grupo de sintéticas (simbolizadas pelas barras verticais mais próximas da cor verde ou azul) que o método gera;

3. No cabeçalho da figura podemos verificar quantas publicações foram geradas usando relevâncias tópicas de reforço e quantas foram geradas usando relevâncias tópicas

de diversificação;

4. Na base horizontal de cada figura representativa do estrato (eixo dos x) estão os números de tópicos, definidos e explicados ao falarmos da análise de tópicos prévia. Ex.: Em Capturing-Creating usamos 22 tópicos, identificados por cada barra colorida dentro da figura Capturing-Creating;

5. A bolinha no centro da barra vertical significa a média da relevância tópica desse tópico no conjunto de publicações, o verde é no conjunto de publicações sintéticas e o laranja é no conjunto de publicações autênticas.

Sendo assim, podemos notar que segundo a Figura 5.1, explicada acima, as publicações autênticas e sintéticas são muito parecidas em relação a média ponderada pela quantidade de publicações que usam aquela relevância tópica.

Publicações autênticas: Cada relevância tópica é usada por exatamente uma publicação autêntica.

Publicações Sintéticas: Uma mesma relevância tópica pode gerar mais de uma publicação sintética de acordo com a *linha 10 do Algoritmo* onde: Para cada publicação sintética é sorteada qual vai ser a sua relevância tópica, então pode ser sorteada uma mesma relevância tópica mais de uma vez. Ou melhor, se a quantidade de publicações sintéticas for maior que a quantidade de publicações autênticas necessariamente terá que haver repetições de relevâncias tópicas, então a média ponderada das relevâncias tópicas sintéticas é relativa a essa quantidade.

Dado que as relevância tópicas de reforço são compartilhadas entre as publicações sintéticas, então, se tivéssemos apenas as relevâncias tópicas de reforço, os 2 quadros da figura iriam ficar exatamente iguais. Porém, também temos as relevâncias tópicas de diversificação que são relevâncias tópicas novas, não usadas nas publicações autênticas, o que nos dá uma pequena variação no gráfico, pequena, como o esperado.

Em alguns pouquíssimos casos, há uma pequena variação da média entre as publicações sejam nas sintéticas, ou nas autênticas, tanto para maior, quanto para menor, porém, as médias entre elas estão quase emparelhadas, demonstrando as similaridades entre as publicações autênticas e sintéticas.

A pequena variação das médias entre as relevâncias tópicas (entre as autênticas e as sintéticas), já era prevista, esses intervalos em torno da média, são os chamados desvio padrão, que atua tanto para mais, quanto para menos.

Ainda precisamos lembrar que duas publicações geradas com a mesma relevância tópica não serão necessariamente iguais. Isto porque a relevância tópica define apenas se elas usarão o mesmo dado de tópico, ou não, porém, para cada palavra o dado de tópico terá que ser jogado novamente, então, sempre haverá a oportunidade de ser sorteado um tópico diferente a cada vez. Sendo assim, por serem usados o mesmo dado de relevância tópica para dois textos distintos, não quer dizer que os textos gerados vão ser iguais. Isto porque ao gerar várias publicações sintéticas com a mesma relevância tópica, as publicações não ficam iguais entre elas, mas elas terão perfis mais aderentes um ao outro.

Na Tabela 5.1 podemos perceber que as correlações entre a relevâncias tópicas autênticas $\boldsymbol{\theta}_a$ e sintéticas $\boldsymbol{\theta}_s$ estão consideravelmente boas. Lembramos que esse comportamento demonstrado até o momento não é sobre as publicações sintéticas que, de fato, usare-

Tabela 5.1: Correlação entre as médias de $\boldsymbol{\theta}_a$ e $\boldsymbol{\theta}_s$.

Estrato	**Correlação**
Capturing-Creating	0.4850532
Disseminating-Storing	0.5782407
Enriching	0.6827514
Interpreting	0.8034473

mos para classificar, esse é um experimento a parte que nos permite verificar como é o comportamento geral da geração de publicações sintéticas.

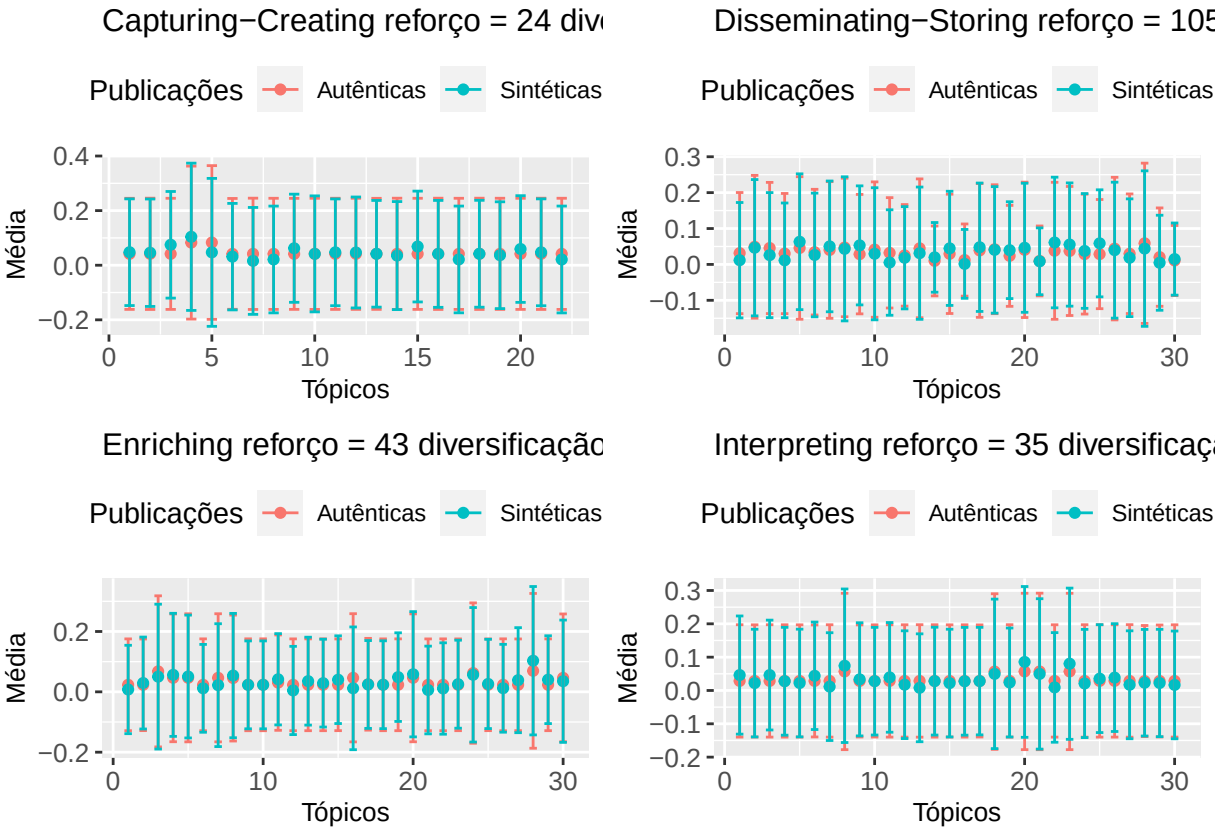


Figura 5.1: Parâmetro $\boldsymbol{\theta}$ médio por estrato e por tópico na geração das publicações sintéticas.

A semelhança entre as publicações sintéticas e autênticas pode ser medida através de uma medida estatística chamada *centralidade* onde comparamos as frequências dos elementos léxicos observadas em relação as esperadas (MANNING; SCHÜTZE, 1999). Ou seja, buscamos expressar a significância das frequências de ocorrências nos corpora de publicações (sintéticas e autênticas). No nosso caso, as frequências observadas são a dos elementos léxicos presentes nas publicações sintéticas S_i em relação aos mesmos elementos léxicos presentes nas publicações autênticas A_i.

Isto porque os elementos léxicos com frequências muito semelhantes em diversas publicações não são bons para identificar a publicação a que pertencem, dado que ele existe em abundância em mais de uma publicação possível. Enquanto que os elementos léxicos de frequência distinta nas publicações, são bons para nos ajudar a distinguirmos as publica-

ções, dado que eles possuem as características fundamentais distintivas dessas publicações presentes nos corpora.

O método que adotamos é o *teste de* χ^2 para cada termo dos corpora, definido como

$$\chi^2 = \sum_{i=1}^{D_a} \frac{(S_i - A_i)^2}{A_i},$$

onde pegamos as diferenças entre os elementos léxicos das publicações sintéticas S_i e das autênticas A_i, elevado ao quadrado para eliminar os números negativos, ponderamos por A_i, tudo isso em somatório de todas as publicações, é equivalente ao quanto essa medida de elementos léxicos será a mais ou a menos do sintético com relação ao autêntico.

Como amostra de resultados dessa medida, apresentamos as figuras 5.2 e 5.3, onde são indicados somente os termos mais representativos, suprimindo os demais pouco distintivos das publicações. Sendo assim:

1. Os valores em azul são positivos e indicam os termos mais presentes, ou que mais identificam as publicações sintéticas, comparativamente as autênticas;
2. Por outro lado, os valores em cinza são negativos indicando os termos mais presentes nas publicações autênticas, ou menos presentes nas sintéticas ou ainda que mais distinguem as autênticas comparativamente às sintéticas.

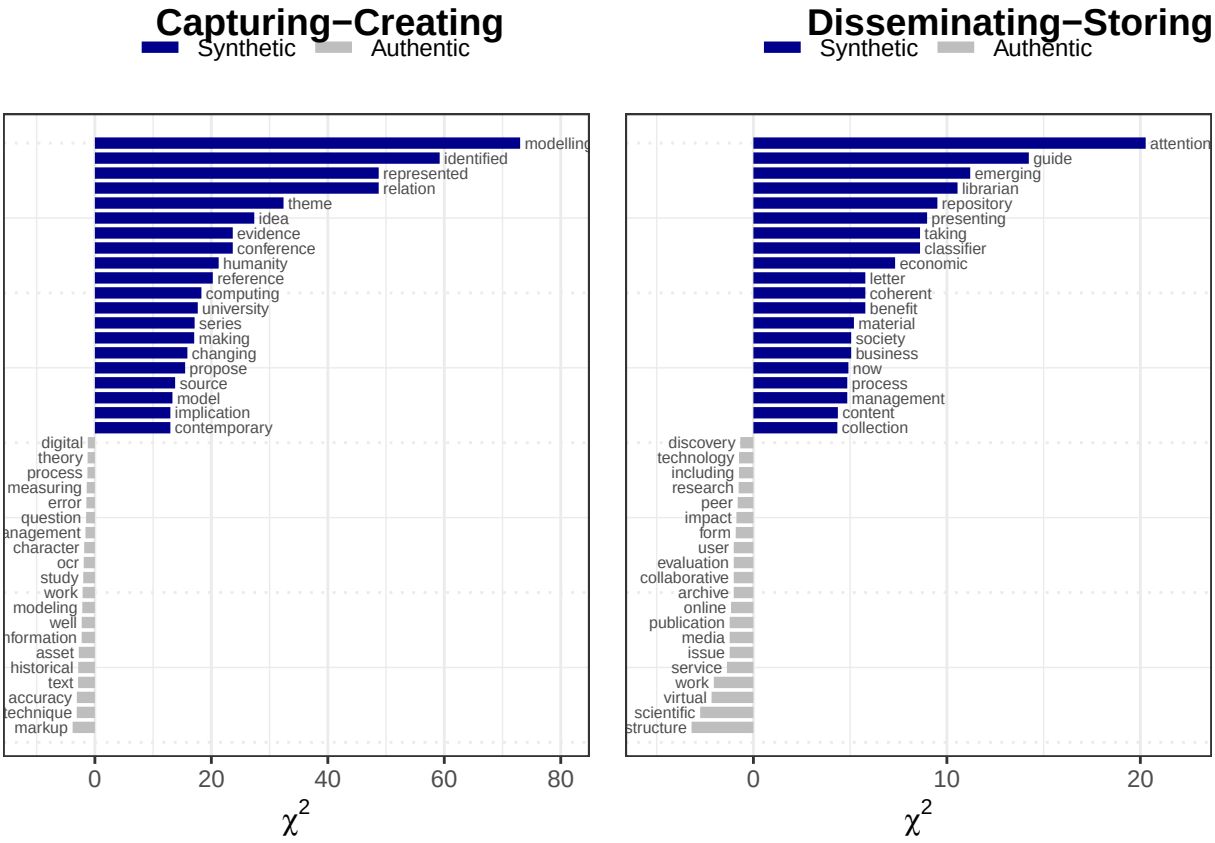


Figura 5.2: Centralidade dos termos nas publicações autênticas e sintéticas.

A verificação visual das figuras 5.2 e 5.3 nos remete a lembrança de que são exibidos somente os termos, que chamamos de elementos léxicos, que distinguem as publicações

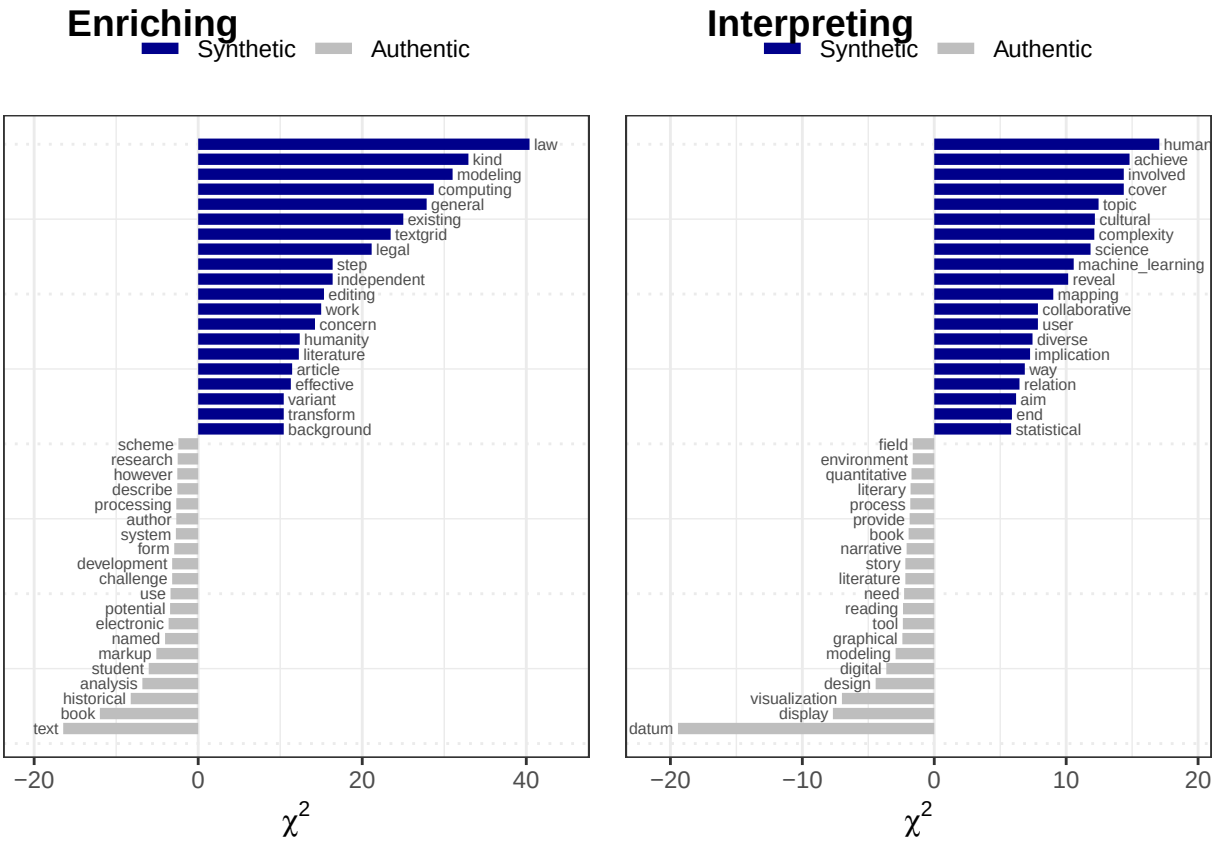


Figura 5.3: Centralidade dos termos nas publicações autênticas e sintéticas.

de forma que se queremos ter textos com as temáticas específicas caracterizadas pelos elementos léxicos que compõem os tópicos associados aos estratos, logo, esses termos não devem aparecer nesta figura, sob pena do experimento não ter cumprido a sua finalidade de replicar o perfil das publicações.

O que significa dizer que, para que tenhamos sucesso em nosso experimento, não podemos ter nas figuras 5.2 e 5.3 elementos léxicos, que determinam o estrato, porém a temática geral deles precisam ser harmônica com cada estrato, pois queremos que as sintéticas representem bem cada estrato.

Sendo assim, após exame visual, podemos afirmar que, observamos raros elementos léxicos nesta condição dos 40 elementos de cada estrato observado, notamos 2 em Enriching, 1 em Interpreting, 4 em Disseminating Storing e nenhum em Capturing-Creating, sendo a temática geral dos elementos léxicos harmônica com o contexto de cada estrato. Portanto consideramos bem sucedida a temática dos elementos léxicos segundo o perfil de cada estrato observado.

5.4 Classificação Bayesiana

Nesta seção, aplicamos os métodos descritos no Capítulo 4, concomitantemente à geração de sintéticas abordada neste Capítulo 5.

Começaremos nossa aplicação real com o LDA de BLEI; NG; JORDAN (2003), passando ao STM de ROBERTS; STEWART; TINGLEY (2019) e por fim, na seção seguinte, analisamos comparativamente os nossos resultados com os do sLDA de BLEI; MCAULIFFE

(2010) que será usado apenas como referência com o objetivo de demonstrar os resultados do nosso método relacionando-o a um outro já em uso.

Sendo assim, seguimos ao experimento, composto de duas etapas clássicas conhecidas como treino e teste.

Corpus: Começamos com o corpus composto pelo total de todas as publicações autênticas **D**. As publicações assim escolhidas, acrescidas dos textos descritivos, formam o corpus para a determinação das relevâncias tópicas através de análise por tópicos via LDA e das relevâncias tópicas. Então seguimos os seguintes passos:

1. Primeiramente separamos aleatoriamente 75% desse corpus que usaremos para o treino, e os outros 25% que usaremos para o nosso teste.
2. Após feito esse sorteio e separação do corpus, então dentro destes 75% do treino, é que fazemos o processo de geração das publicações sintéticas;
3. Agora que as publicações sintéticas já foram geradas, então as incluímos ao corpus, onde obtemos um *corpus com publicações autênticas e sintéticas*;
4. Só então é que passamos a aplicação do nosso método a esse conjunto de treino composto pela soma de 75% das publicações autênticas com as sintéticas.

Aplicação do modelo:

5. Aplicamos o nosso método para gerar o modelo no *corpus com as publicações autênticas e sintéticas* que separamos para o treino;
6. Depois, pegamos o modelo gerado e aplicamos nos outros 25% das publicações autênticas restantes que separamos para o teste, ou seja, para a classificação segundo (4.1).

Sendo assim, a aplicação é parecida com o método do Capítulo 4, a única diferença é que agora agregamos as publicações sintéticas na fase de treino.

7. Então, repetimos esse treino e teste 10 vezes. Então, para cada experimento (treino e teste) geramos as sintéticas daquele experimento, só com as publicações autênticas de treino.

Com isso, geramos o modelo do nosso método e fazemos a classificação das publicações autênticas de teste daquele experimento. Repetindo esse processo 10 vezes, ou seja, fazendo 10 experimentos.

8. por último, pegamos a média do comportamento para traçar as tabelas que mostramos aqui nos resultados do LDA e do STM. Que são novas versões das tabelas do Capítulo 4, usando as publicações sintéticas.

5.4.1 Relevância Tópica via LDA

Aplicando o método bayesiano (4.1) tendo as relevâncias tópicas determinadas por meio da análise por tópicos com LDA, só que agora usando as publicações sintéticas. Esse processo é repetido 10 vezes, e as médias das avaliações de cada repetição são mostradas na Tabela 5.2, onde podemos notar uma melhora significativa dos resultados, principalmente nos estratos desbalanceados.

Porém, os resultados mais precisos ainda são fixados nos estratos mais fortemente representados por tópicos (Analyzing e Disseminating-Storing), como era de se esperar, já que a determinação do estrato para cada tópico define os temas que serão tratados em cada estrato ao passo que o desbalanceamento aprofunda a representatividade de determinados tópicos em alguns estratos. Ou seja, os efeitos do desbalanceamento da amostra continua atuante, ainda que atenuados devido a presença das publicações sintéticas no corpus do experimento.

Tabela 5.2: Atribuição das publicações aos estratos com o método (4.1) e LDA, incluindo amostras sintéticas (média sobre 10 amostragens).

Estrato	**Precisão**	**Recolhimento**	**Pontuação**
Analyzing	0.8387948	0.8745921	0.8557538
Capturing-Creating	0.5150000	0.5666667	0.5321429
Disseminating-Storing	0.7661390	0.8264022	0.7915910
Enriching	0.6353175	0.3557937	0.4429382
Interpreting	0.5079365	0.4154762	0.4388545

5.4.2 Relevância Tópica via STM

O desempenho do STM foi similar ao do LDA com uma ligeira melhora dos valores, provavelmente atribuído às mesmas correlações que influenciaram uma também ligeira melhora do STM em relação ao LDA no Capítulo 4. Os resultados da aplicação de (4.1) com STM estão demonstrados na Tabela 5.3.

Isto acontece porque o STM é dependente da existência de metadados relevantes associados às publicações, o que não existe em nosso experimento com as sintéticas, já que uniformizamos os metadados das publicações sintéticas que geramos, ou seja, empregamos os mesmos metadados utilizados no Capítulo 4, a saber $\langle$`ItemType`, `PublicationYear`, `Class`$\rangle$ também neste capítulo, com o objetivo de uniformizar o experimento.

Porém, insistimos que, caso haja metadados disponíveis no conjunto de publicações, então este experimento via STM tende a ficar ainda melhor. O que não aconteceu no nosso caso, já que a forma de geração dos metadados das amostras sintéticas que utilizamos é constituída de um só grupo de valores dos metadados para todos os estratos.

Acreditamos que se o conjunto de metadados fossem separados, provavelmente teríamos um resultado ainda melhor, porém, mesmo assim, o resultado do STM ainda foi melhor que o do LDA em todos os estratos, ainda que não tanto quanto esperávamos.

5.5 Resultados: LDA Supervisionado

No método LDA supervisionado binário que usamos para avaliar o nosso método, os experimentos tiveram um grande impacto negativo devido as publicações sintéticas usadas para contornar o desbalanceamento dos estratos. Cada estrato é classificado binariamente em relação ao somatório dos outros estratos, o que equivale a dizer que o desbalanceamento

Tabela 5.3: Atribuição das publicações aos estratos com o método (4.1) e STM, incluindo amostras sintéticas (média sobre 10 amostragens).

Estrato	Precisão	Recolhimento	Pontuação
Analyzing	0.8935645	0.8308256	0.8598193
Capturing-Creating	0.6847619	0.5680952	0.6098590
Disseminating-Storing	0.7719481	0.8835935	0.8221636
Enriching	0.5885634	0.4646032	0.5032693
Interpreting	0.3593506	0.5226190	0.4096717

é aprofundado ainda mais para todos os estratos, mesmo o majoritário *Analyzing*.

A sequência de experimentos na classificação binária, com a consequente escolha das publicações sintéticas, é feita de forma aleatória e em proporções variadas onde para cada estrato i, são realizados 10 experimentos de classificação sendo que, ao final, cada publicação autêntica é associada ao estrato em que ela mais foi classificada.

Sendo assim, concluímos que, tendo em vista que usamos o LDA Supervisionado como forma de comparação com o método que propomos, notamos que o método bayseano demonstrou resultados satisfatórios, para o nosso contexto, dado os desafios do desbalanceamento e das sobreposições nas amostras.

No Capítulo 4 já notamos este fato, incluindo os melhores resultados do nosso método em relação ao sLDA, devido às sobreposições de estratos, porém piorou ainda mais os resultados neste Capítulo, dado o aprofundamento do desbalanceamento provocado pela classificação binária usando as publicações geradas sinteticamente. Ou seja, o desbalanceamento e principalmente a sobreposição de estratos, prejudicam o uso do LDA Supervisionado. Sendo que o desbalanceamento prejudica a classificação e a sobreposição inviabiliza o uso não binário do sLDA.

Tabela 5.4: Atribuição das publicações aos estratos com o método (4.1) e sLDA, incluindo amostras sintéticas (média sobre 10 amostragens).

Estrato	Precisão	Recolhimento	Pontuação
Analyzing	0.5114435	0.2682212	0.3267125
Capturing-Creating	0.0570833	0.1261905	0.0747826
Disseminating-Storing	0.5060213	0.2727891	0.2973141
Enriching	0.1333250	0.3837771	0.1873763
Interpreting	0.3781300	0.5679762	0.3930663

6 CONCLUSÕES

A quantidade de informações disponíveis hoje em dia é tão grande que é necessário criar ferramentas para organizar e representar essas informações de forma que elas possam ser mais facilmente detectáveis, pesquisáveis e acessadas. Pesquisadores do campo das Humanidades Digitais desenvolveram a *TaDiRAH*, um sistema de representação e organização da informação para atender as demandas interdisciplinares próprias de um campo marcado pela intercessão da Computação com as Ciências Humanas. Sua viabilidade relaciona-se ao uso prático e hierárquico de uma taxonomia (mais adequada para ambientes digitais).

Para criar a *TaDiRAH*, foi feito um consórcio de instituições que já tinham expertises em catalogar, organizar e representar recursos usados em Humanidades Digitais, e criaram uma taxonomia onde o caráter prático é o elemento essencial desde a sua elaboração até o seu uso. Criada para ser uma ferramenta útil, necessária e agradável aos praticantes de Humanidades Digitais, por isso ela foi dividida para ser usada livremente de acordo com os interesses particulares de cada usuário e ao mesmo tempo poder ser uma ferramenta de desenvolvimento, integração e auxílio aos pesquisadores e instituições na busca e compartilhamento de recursos de pesquisas.

A *TaDiRAH* é uma ferramenta de, com e para humanistas digitais, e está em perfeita sintonia com o Manifesto das Humanidades Digitais criado no THATCamp 2010, essa singularidade a torna uma importante peça no desenvolvimento e compartilhamento de recursos importantes para o campo, além de ser um auxílio importante para o pensar, entender e desenvolver o campo. Sendo assim, entendemos ser importante a criação e o desenvolvimento de ferramentas para: (1) Auxiliar no entendimento do campo e suas dinâmicas através de uma lente comum ao próprio campo; (2) Auxiliar no uso desta taxonomia, dado que ninguém é obrigado a usá-la em suas pesquisas e compartilhamento de informações, mas, com certeza ela é uma ferramenta útil ao campo das Humanidades Digitais; (3) Colaborar para o pensar soluções de apoio a otimização do uso da taxonomia.

Ao analisarmos a *TaDiRAH* podemos perceber que sua estrutura não disjunta procura atender a demandas de trabalhos de campos distintos com expertises e práticas que nem sempre podem ser nitidamente separadas. O que provoca sobreposição de estratos dentro da própria estrutura hierárquica da taxonomia a partir do terceiro nível de seus substratos, ou seja, alguns substratos participam de mais de um estrato de nível superior.

As sobreposições nos estratos da *TaDiRAH*, ao nosso ver, é uma tentativa de atender a demandas relacionadas a imbricamentos próprios de três conjuntos de atividades, objetos e métodos de pesquisas, dos pesquisadores de Humanidades Digitais: (1) as demandas mais voltados para as Ciências Humanas positivadas nos estratos *Analyzing*, *Enriching* e *Interpreting*; (2) As demandas de áreas mais computacionais e técnicas dos elementos digitais do campo, representadas pelos estratos *Capturing* e *Creating* e (3) as demandas gerais mais transversalizadas pela necessidade cada vez mais crescente de digitalizar processos (produtos e serviços) no sentido de finalizar, distribuir e disponibilizar, para o público em geral, os objetos digitais e publicações produzidos pelos pesquisadores.

Notamos também que cada nível e subnível dos estratos acompanha um texto descritivo que o explica, além do próprio nome do estrato ou substrato que pode ser usado como palavra-chave, o que significa dizer que os estratos de primeiro nível que possuem mais

substratos, e por consequência, também possuem mais textos descritivos que explicam o próprio estrato, possuem mais facilidade de uso tanto por seres humanos quanto em buscas automatizadas em bancos de dados. O que pode explicar o fato do extrato com mais subníveis (*Analyzing*) possuir muito mais textos classificados do que os demais estratos.

Dado o nosso objetivo primário de produzir uma ferramenta de classificação automatizada usando a taxonomia *TaDiRAH*, justamente por ser esta, um sistema completo que busca representar e explicar o campo das Humanidades Digitais. Considerando também o fato de que as ferramentas de análise automatizadas de textos são essenciais na organização e recuperar informações em grande corpora de textos, na pesquisa de padrões cujos olhos humanos não conseguiriam identificar dado a enorme quantidade de textos que se tem para analisar e na observação de que as ferramentas computacionais podem ser manipuladas para atender a demandas distintas, de acordo com os necessidade da nossa pesquisa.

Examinamos diversas abordagens, encontrando mais sinergia de nossos objetivos com os modelos de: Atribuição de Tópicos por *Dirichlet* (*LDA*) (BLEI; NG; JORDAN, 2003), Atribuição de Tópicos por Correlação das Relevâncias (*CTM*) (BLEI; LAFFERTY, 2006b), Atribuição de Tópicos por Estrutura (*STM*) (ROBERTS et al., 2013) e Atribuição por Distribuição Latente de Dirichlet Supervisionada (*sLDA*) (BLEI; MCAULIFFE, 2010). Porém descartamos o uso do *CTM* devido às dificuldades técnicas envolvidas na complexidade dos cálculos que exigem muito da máquina.

Para realizar o experimento, precisamos coletar dois conjuntos de publicações de Humanidades Digitais, um já classificado por especialistas da área e outro ainda não classificado; preparar os dados para as análises, limpando e organizando os textos para separar as palavras que realmente interessa; analisar preliminarmente os dados, onde constatamos algumas dificuldades no primeiro conjunto (treino e teste) como as sobreposições e os desbalanceamentos de estratos nos textos, como já havíamos previsto durante a análise da estrutura da *TaDiRAH*.

Para solucionar essas dificuldades, testamos diversas soluções. A primeira dificuldade relaciona-se ao fato de haver muitas publicações classificadas em mais de um estrato da *TaDiRAH*, são as chamadas sobreposições de estrato. Para lidar com essa demanda desenvolvemos um método Bayesiano que classifica os documentos usando o perfil de construção das publicações mapeada pelo modelo de análise por tópico com a relevância de cada tópico em cada estrato dada pelo cálculo da pertinência tópica que fazemos usando o *LDA* e o *STM*.

Os resultados encontrados foram melhores dos que os que já tínhamos usando cada método separadamente. Usamos então o *sLDA* como padrão comparativo com o nosso método, já que ele é um método em uso pela comunidade de pesquisadores. Fizemos então uma classificação binária, comparando um único estrato por vez contra o somatório de todos os outros restantes, pois o método *sLDA* usa a regressão logística para classificar as publicações, sendo que uma regressão pressupõe que não pode haver sobreposição de estratos. O nosso método obteve resultados melhores que o *sLDA*.

Também percebemos que poderíamos melhorar ainda mais os resultados se conseguíssemos diminuir o desbalanceamento entre os estratos, pois originalmente temos um desbalanceamento com peso de 3 a 9 vezes do extrato majoritário *Analyzing* em relação aos demais. Buscamos diversas soluções, mas todas eram voltadas para dados numéricos, então resolvemos usar o modelo com o perfil das publicações construído pela análise de tópicos para

produzir textos sintéticos com o intuito de combater o desbalanceamento e fazer novas análises.

Notamos uma melhora significativa dos resultados no *LDA*, principalmente nos estrados minoritários, porém os resultados mais precisos ainda ficaram com os estratos *Analyzing* e *Disseminating-Storing*. Os resultados usando o *STM* foram ainda mais satisfatórios que os do *LDA*, mas esperávamos um resultado ainda melhor, que não foram alcançados devido os metadados dos textos sintéticos que foram todos iguais, e por isso irrelevantes para a análise. Comparativamente ao *sLDA*, os textos sintéticos obtiveram um grande impacto negativo na análise piorando ainda mais os seus resultados, isto porque como tivemos mais publicações nos demais estratos, aumentamos o desbalanceamento da classificação binária, ao classificar cada estrato com o somatório dos demais.

Sendo assim, tendo em vista os resultados apresentados nos experimentos realizados, concluímos que o classificador Bayesiano obteve resultados satisfatórios dado o contexto do nosso trabalho com os desafios enfrentados de sobreposição e desbalanceamento na estrutura da *TaDiRAH* e nos documentos classificados artesanalmente.

Devido a interseção entre os estratos, consideramos que os termos da *TaDiRAH* podem não ser disjuntos o suficiente para classificar os textos dentro de apenas uma classe, sendo assim, e enquanto esta realidade se manifestar, pode ser que não consigamos uma acurácia ideal, porém, cremos que poderemos ficar satisfeitos se o algoritmo conseguir classificar as publicações em qualquer um dos estratos nos quais a publicação estiver classificada.

Também observamos uma tendência do campo em produzir publicações mais voltadas para o estrato *Analyzing*. Esse comportamento nos parece ser coerente com o fato deste estrato possuir mais substratos que os demais, e consequentemente, mais textos descritivos. Corrobora também para esta conclusão o Grupo do Zotero da DARIAH, de onde obtivemos as publicações previamente classificadas por especialistas de Humanidades Digitais, manter unidos em um único grupo os estratos imbricados *Disseminating_Storing* e *Capturing-Creating*.

Por fim, sugerimos como trabalhos futuros o emprego adicional de outros métodos de tratamento de desbalanceamento. Os números de acurácia, precisão e recolhimento, também poderão ser aprimorados em trabalhos futuros, ajustando o algoritmo, obtendo mais publicações classificadas manualmente por especialistas de Humanidades Digitais, ou através de alguma outra oportunidade de melhoria que possa ser visualizada em novas oportunidades de revisões.

REFERÊNCIAS BIBLIOGRÁFICAS

AGARWAL, D.; CHEN, B.-C. **fLDA: matrix factorization through latent dirichlet allocation**Proceedings of the third ACM international conference on Web search and data mining. **Anais**...: WSDM '10.New York, NY, USA: Association for Computing Machinery, fev. 2010. Acesso em: 25 ago. 2021

AHMED, A.; XING, E. **Staying Informed: Supervised and Semi-Supervised Multi-View Topical Analysis of Ideological Perspective**Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. **Anais**...Cambridge, MA: Association for Computational Linguistics, out. 2010. Acesso em: 23 set. 2021

ALGHAMDI, R.; ALFALQI, K. A Survey of Topic Modeling in Text Mining. **International Journal of Advanced Computer Science and Applications**, v. 6, n. 1, 2015.

ALTMAN, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. **The American Statistician**, v. 46, n. 3, p. 175–185, 1992.

BAI, Y.; WANG, J. **News Classifications with Labeled LDA:**Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. **Anais**...Lisbon, Portugal: SCITEPRESS - Science; and Technology Publications, 2015. Acesso em: 14 out. 2021

BAO, S. et al. Joint Emotion-Topic Modeling for Social Affective Text Mining. **2009 Ninth IEEE International Conference on Data Mining**, 2009.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20–29, jun. 2004.

BERGHOLZ, A. et al. **Improved phishing detection using model-based features**In Fifth Conference on Email and Anti-Spam, CEAS. **Anais**...2008

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, v. 55, n. 4, p. 77–84, abr. 2012.

BLEI, D. M.; GRIFFITHS, T. L.; JORDAN, M. I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. **Journal of the ACM**, v. 57, n. 2, p. 7:1–7:30, fev. 2010.

BLEI, D. M.; LAFFERTY, J. D. **Dynamic topic models**Proceedings of the 23rd international conference on Machine learning - ICML '06. **Anais**...Pittsburgh, Pennsylvania: ACM Press, a2006. Acesso em: 21 ago. 2021

BLEI, D. M.; LAFFERTY, J. D. A correlated topic model of Science. **The Annals of Applied Statistics**, v. 1, n. 1, p. 17–35, jun. 2007.

BLEI, D. M.; LAFFERTY, J. D. **Topic Models**Taylor; Francis Group, LLC, 2009. Acesso em: 9 out. 2021

BLEI, D. M.; MCAULIFFE, J. D. **Supervised Topic Models**Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., dez. 2007. Acesso em: 28 abr. 2021

BLEI, D. M.; MCAULIFFE, J. D. Supervised Topic Models. **arXiv:1003.0783 [stat]**, mar. 2010.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. p. 30, 2003.

BLEI, D.; LAFFERTY, J. D. **Correlated Topic Models**Advances in Neural Information Processing Systems. **Anais**...MIT Press, b2006. Acesso em: 6 set. 2021

BLOCH, M. Apologia da História ou O Ofício de Historiador. p. 153, 2002.

BOREK, L. et al. **TOWARDS A PRACTICAL TAXONOMY OF DIGITAL HUMANITIES RESEARCH ACTIVITIES AND OBJECTSDigial Humanities - Lausanne - Switzeland '14**, jul. 2014. Acesso em: 15 maio. 2020

BOREK, L. et al. TaDiRAH: a Case Study in Pragmatic Classification. **Digital Humanities Quarterly**, v. 010, n. 1, fev. 2016.

BOREK, L. et al. **TaDiRAH: Taxonomy of Digital Research Activities in the Humanities**, set. 2020. Acesso em: 23 jul. 2021

BOREK, L. et al. **TaDiRAH v2.0.1**, jul. 2021. Disponível em: <<https://vocabs.dariah.eu/tadirah/en/>>. Acesso em: 21 out. 2021

BRAMER, M. Ensemble Classification. Em: **Principles of Data Mining**. [s.l.] Springer, 2013.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, out. 2001.

CAO, J. et al. A density-based method for adaptive LDA model selection. **Neurocomputing**, Advances em Machine Learning e Computational Intelligence. v. 72, n. 7, p. 1775–1781, mar. 2009.

CASTRO, R. M. DE. **Análise da literatura das humanidades digitais: uma proposta bibliométrica para descrição de seu escopo e congruência conceitual**. tese de doutorado—[s.l: s.n.].

CHAMORRO, W. et al. Listado de especies y clave de generos y subgeneros de escarabajos estercoleros (Coleoptera: Scarabaeidae: Scarabaeinae) presentes y presuntos para Ecuador. **Revista Colombiana de Entomologia**, v. 44, n. 1, p. 72–101, jan. 2018.

CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, jun. 2002.

CHIENA, J.-T.; LEEA, C.-H.; TANB, Z.-H. **Latent Dirichlet mixture model - ScienceDirect**, set. 2017. Acesso em: 25 ago. 2021

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995.

COSTA, G.; ORTALE, R. Jointly modeling and simultaneously discovering topics and clusters in text corpora using word vectors. **Information sciences**, v. 563, p. 226–240, 2021.

DACOS, M. **Manifesto das digital humanitiesTHATCamp Paris**, 2011. Acesso em: 7 fev. 2021

DARIAH-UE CONSORTIUM. **Zotero | Groups > Doing Digital Humanities - A DARIAH Bibliography**, nov. 2012. Acesso em: 9 mar. 2021

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American Society for Information Science**, v. 41, n. 6, p. 391–407, 1990.

DHTAXONOMY. **TaDiRAH (github)**Digital Humanities Taxonomy Group, 2020. Acesso em: 15 maio. 2020

EBENUWA, S. H. **Handling Imbalanced Classes: Feature Based Variance Ranking Techniques for Classification**. text—[s.l.] University of East London, set. 2019.

EISENSTEIN, J. et al. **A Latent Variable Model for Geographic Lexical Variation**Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. **Anais**...Cambridge, MA: Association for Computational Linguistics, out. 2010. Acesso em: 23 set. 2021

EISENSTEIN, J.; AHMED, A.; XING, E. **Sparse Additive Generative Models of Text**ICML. **Anais**...2011

FORTES, A.; ALVIM, L. Evidências, códigos e classificações: o ofício do historiador e o mundo digital. **Esboços: histórias em contextos globais**, v. 27, p. 207–227, jun. 2020.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. **Proceedings of the National Academy of Sciences**, v. 101, n. suppl 1, p. 5228–5235, abr. 2004.

GRIMMER, J.; STEWART, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. **Political Analysis**, v. 21, n. 3, p. 267–297, 2013.

GROSSETTI, F. **OpTop: detect the optimal number of topics from a pool of LDA models**, jul. 2021. Disponível em: <<https://github.com/contefranz/OpTop>>. Acesso em: 27 out. 2021

GROSSETTI, F.; LEWIS, C. **OpTop: detect the optimal number of topics from a pool of LDA models**. [s.l: s.n.].

HINGMIRE, S. et al. **Document Classification by Topic Labeling**. [s.l: s.n.].

HOFMANN, T. Probabilistic Latent Semantic Indexing. **ACM SIGIR Forum**, v. 51, n. 2, p. 211–218, ago. 2017.

HONG, L.; DAVISON, B. D. **Empirical study of topic modeling in Twitter**Proceedings of the First Workshop on Social Media Analytics. **Anais**...: SOMA '10.New York, NY, USA: Association for Computing Machinery, jul. 2010. Acesso em: 20 ago. 2021

HOSPEDALES, T.; GONG, S.; XIANG, T. Video Behaviour Mining Using a Dynamic Topic Model. **International Journal of Computer Vision**, v. 98, n. 3, p. 303–323, jul. 2012.

JAPKOWICZ, N. **The Class Imbalance Problem: Significance and Strategies**In

Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI. **Anais**...2000

KHALIFA, O. et al. Multi-objective Topic Modeling. Em: HUTCHISON, D. et al. (Eds.). **Evolutionary Multi-Criterion Optimization**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. v. 7811p. 51–65.

KOZIARSKI, M. Radial-Based Undersampling for imbalanced data classification. **Pattern Recognition**, v. 102, p. 107262, 2020.

KOZIARSKI, M.; KRAWCZYK, B.; WOŹNIAK, M. Radial-Based Approach to Imbalanced Data Oversampling. Em: MARTÍNEZ DE PISÓN, F. et al. (Eds.). **Hybrid Artificial Intelligent Systems – HAIS 2017**. Lecture Notes em Computer Science. [s.l.] Springer, 2017. v. 10334.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. **Progress in Artificial Intelligence**, v. 5, n. 4, p. 221–232, nov. 2016.

LEWIS, C.; GROSSETTI, F. **A Statistical Approach for Optimal Topic Model Identification**. [s.l: s.n.].

LING, C. X.; SHENG, V. S. Cost-Sensitive Learning. Em: SAMMUT, C.; WEBB, G. I. (Eds.). **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 231–235.

LIU, L. et al. An overview of topic modeling and its current applications in bioinformatics. **SpringerPlus**, v. 5, n. 1, p. 1608, set. 2016.

LUZ, F. F. **Consulta a ontologias utilizando linguagem natural controlada**. text— [s.l.] Universidade de São Paulo, out. 2013.

MACHADO, F. S. Scientific Divulgation and Digital Utterances. **Bakhtiniana: Revista de Estudos do Discurso**, v. 11, p. 93–110, ago. 2016.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. [s.l.] The MIT Press, 1999.

MCCALLUM, A.; WANG, X.; CORRADA-EMMANUEL, A. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. **Journal of Artificial Intelligence Research**, v. 30, p. 249–272, out. 2007.

MEYER, E. T.; ECCLES, K. **The Impacts of Digital Collections: Early English Books Online & House of Commons Parliamentary Papers**. Rochester, NY: Social Science Research Network, mar. 2016. Acesso em: 29 out. 2021.

MIMNO, D.; MCCALLUM, A. Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. **arXiv:1206.3278 [cs, stat]**, jun. 2012.

PERKINS, J. et al. Building Bridges to the Future of a Distributed Network: From DiRT Categories to TaDiRAH, a Methods Taxonomy for Digital Humanities. **International Conference on Dublin Core and Metadata Applications**, p. 181–183, out. 2014.

RIBEIRO, F. A. D. S. et al. EXPLORANDO OS POTENCIAIS DA HISTÓRIA DIGITAL: A EXPERIÊNCIA DO CENTRO DE DOCUMENTAÇÃO E IMAGEM DA UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO - CAMPUS DE NOVA IGUAÇU. **Estudos Históricos (Rio de Janeiro)**, v. 33, n. 69, p. 152–172, abr. 2020.

ROBERTS, M. E. et al. The Structural Topic Model and Applied Social Science. p. 4, 2013.

ROBERTS, M. E. et al. Structural Topic Models for Open-Ended Survey Responses: STRUCTURAL TOPIC MODELS FOR SURVEY RESPONSES. **American Journal of Political Science**, v. 58, n. 4, p. 1064–1082, out. 2014.

ROBERTS, M. E.; STEWART, B. M.; TINGLEY, D. stm: An R Package for Structural Topic Models. **Journal of Statistical Software**, v. 91, n. 1, p. 1–40, out. 2019.

ROSEN-ZVI, M. et al. The Author-Topic Model for Authors and Documents. **arXiv:1207.4169 [cs, stat]**, p. 487–494, jul. 2012.

SALTON, G. Some research problems in automatic information retrieval. **ACM SIGIR Forum**, v. 17, n. 4, p. 252–263, jun. 1983.

SHEN, Z.-Y.; SUN, J.; SHEN, Y.-D. **Collective Latent Dirichlet Allocation**2008 Eighth IEEE International Conference on Data Mining. **Anais**...Pisa, Italy: IEEE, dez. 2008. Acesso em: 2 set. 2021

SONG, J. et al. **A bi-directional sampling based on K-means method for imbalance text classification**2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). **Anais**...2016

STEYVERS, M.; GRIFFITHS, T. Probabilistic Topic Models. Em: **Handbook of Latent Semantic Analysis**. [s.l.] Psychology Press, 2007.

SUN, Y. et al. Cost-Sensitive Boosting for Classification of Imbalanced Data. **Pattern Recogn.**, v. 40, n. 12, p. 3358–3378, dez. 2007.

TADIRAH. **TaDiRAH - Taxonomy of Digital Research Activities in the Humanitieshttps://vocabularyserver.com/tadirah/en/**, jul. 2014. Acesso em: 8 mar. 2021

TEMPLETON, C. **Topic Modeling in the Humanities: An Overview**, 2011. Acesso em: 7 out. 2021

TERRAS, M. M.; NYHAN, J.; VANHOUTTE, E. (EDS.). **Defining digital humanities: a reader**. Farnham, Surrey, England : Burlington, VT: Ashgate Publishing Limited ; Ashgate Publishing Company, 2013.

THOMPSON, E. P. **The Poverty of Theory: Or an Orrery of Errors**. [s.l.] Merlin Press, 1996.

UNSWORTH, J. What is humanities Computing and What is Not? Em: TERRAS, M. M.; NYHAN, J.; VANHOUTTE, E. (Eds.). Farnham, Surrey, England : Burlington, VT: Ashgate Publishing Limited ; Ashgate Publishing Company, 2013. p. 35–48.

VAYANSKY, I.; KUMAR, S. A. P. A review of topic modeling methods. **Information Systems**, v. 94, p. 101582, dez. 2020.

VIGNOLI, R. G.; SOUTO, D. V. B.; CERVANTES, B. M. N. Sistemas de organização do conhecimento com foco em ontologias e taxonomias. **Informação & Sociedade**, v. 23, n. 2, p. 59–72, jul. 2013.

VIKRAMKUMAR; B, V.; TRILOCHAN. Bayes and Naive Bayes Classifier. **ar-**

Xiv:1404.0933 [**cs**], abr. 2014.

VITAL, L. P.; CAFÉ, L. M. A. Ontologias e taxonomias: diferenças. **Perspectivas em Ciência da Informação**, v. 16, n. 2, p. 115–130, 2011.

VORONTSOV, K.; POTAPENKO, A. **Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization** (D. I. Ignatov et al., Eds.)Analysis of Images, Social Networks and Texts. **Anais**...: Communications em Computer e Information Science.Cham: Springer International Publishing, 2014

WANG, X.; MCCALLUM, A. **Topics over time: a non-Markov continuous-time model of topical trends**Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06. **Anais**...Philadelphia, PA, USA: ACM Press, 2006. Acesso em: 2 set. 2021

WATANABE, K. Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. **Communication Methods and Measures**, v. 0, n. 0, p. 1–22, nov. 2020.

WATANABE, K.; ZHOU, Y. Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches. **Social Science Computer Review**, p. 0894439320907027, fev. 2020.

WELBERS, K.; ATTEVELDT, W. V.; BENOIT, K. Text Analysis in R. **Communication Methods and Measures**, v. 11, n. 4, p. 245–265, out. 2017.

YAU, C.-K. et al. Clustering scientific documents with topic modeling. **Scientometrics**, v. 100, n. 3, p. 767–786, abr. 2014.

ZHAO, W. et al. A heuristic approach to determine an appropriate number of topics in topic modeling. **BMC Bioinformatics**, v. 16, n. 13, p. S8, dez. 2015.

ZHOU, Z.; ZHOU, J.; ZHANG, L. **Demand-adaptive Clothing Image Retrieval Using Hybrid Topic Model**Proceedings of the 24th ACM international conference on Multimedia. **Anais**...: MM '16.New York, NY, USA: Association for Computing Machinery, out. 2016. Acesso em: 21 ago. 2021

ZUO, Y. et al. **Topic Modeling of Short Texts: A Pseudo-Document View**Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**...San Francisco California USA: ACM, ago. 2016. Acesso em: 26 ago. 2021