UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

FELIPE PENHORATE CARVALHO DA FONSECA

Inferência das áreas de atuação de pesquisadores

São Paulo
2018

FELIPE PENHORATE CARVALHO DA FONSECA

Inferência das áreas de atuação de pesquisadores

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 30 de janeiro de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2018

CATALOGAÇÃO-NA-PUBLICAÇÃO
(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Fonseca, Felipe Penhorate Carvalho da
Inferência das áreas de atuação de pesquisadores / Felipe Penhorate Carvalho da Fonseca ; orientador, Luciano Antonio Digiampietri. – 2018.
106 f.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.
Versão corrigida

1. Inteligência artificial. 2. Representação de conhecimento. 3. Redes sociais. 4. Plataforma Lattes. 5. Sistemas baseados em conhecimento. I. Digiampietri, Luciano Antonio, orient. II. Tìtulo.

CDD 22.ed.– 006.3

Dissertação de autoria de Felipe Penhorate Carvalho da Fonseca, sob o título **"Inferência das áreas de atuação de pesquisadores"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 30 de janeiro de 2018 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri

Instituição: Universidade de São Paulo

Presidente

Prof. Dr. Ivandré Paraboni

Instituição: Universidade de São Paulo

Prof. Dr. Pedro Olmo Stancioli Vaz de Melo

Instituição: Universidade Federal de Minas Gerais

Prof. Dr. João Eduardo Ferreira

Instituição: Universidade de São Paulo

Agradecimentos

Agradeço à CAPES por ter fomentado essa pesquisa por meio de uma bolsa de mestrado, a minha família por todo apoio dado durante todo o mestrado, seja financeiro quanto emocional e a todos professores do curso, em especial ao meu orientador Luciano Antonio Digiampietri, por terem me orientado, ajudado e motivado para que esta pesquisa fosse completa.

Resumo

FONSECA, Felipe Penhorate Carvalho. **Inferência das áreas de atuação de pesquisadores**. 2018. 106 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Atualmente, existe uma grande gama de dados acadêmicos disponíveis na web. Com estas informações é possível realizar tarefas como descoberta de especialistas em uma dada área, identificação de potenciais bolsistas de produtividade, sugestão de colaboradores, entre outras diversas. Contudo, o sucesso destas tarefas depende da qualidade dos dados utilizados, pois dados incorretos ou incompletos tendem a prejudicar o desempenho dos algoritmos aplicados. Diversos repositórios de dados acadêmicos não contêm ou não exigem a informação explícita das áreas de atuação dos pesquisadores. Nos dados dos currículos Lattes essa informação existe, porém é inserida manualmente pelo pesquisador sem que haja nenhum tipo de validação (e potencialmente possui informações desatualizadas, faltantes ou mesmo incorretas). O presente trabalho utilizou técnicas de aprendizado de máquina na inferência das áreas de atuação de pesquisadores com base nos dados cadastrados na plataforma Lattes. Os títulos da produção científica foram utilizados como fonte de dados, sendo estes enriquecidos com informações semanticamente relacionadas presentes em outras bases, além de adotar representações diversas para o texto dos títulos e outras informações acadêmicas como orientações e projetos de pesquisa. Objetivou-se avaliar se o enriquecimento dos dados melhora o desempenho dos algoritmos de classificação testados, além de analisar a contribuição de fatores como métricas de redes sociais, idioma dos títulos e a própria estrutura hierárquica das áreas de atuação no desempenho dos algoritmos. A técnica proposta pode ser aplicada a diferentes dados acadêmicos (não sendo restrita a dados presentes na plataforma Lattes), mas os dados oriundos dessa plataforma foram utilizados para os testes e validações da solução proposta. Como resultado, identificou-se que a técnica utilizada para realizar o enriquecimento do texto não auxiliou na melhoria da precisão da inferência. Todavia, as métricas de redes sociais e representações numéricas melhoram a inferência quando comparadas com técnicas do estado da arte, assim como o uso da própria estrutura hierárquica de classes, que retornou os melhores resultados dentre os obtidos.

Palavras-chaves: Classificação de texto. Enriquecimento de texto. Inferência de áreas de atuação. Modelagem de tópicos. Plataforma Lattes.

Abstract

FONSECA, Felipe Penhorate Carvalho. **Inference of the area of expertise of researchers**. 2018. 106 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Nowadays, there is a wide range of academic data available on the web. With this information, it is possible to solve tasks such as the discovery of specialists in a given area, identification of potential scholarship holders, suggestion of collaborators, among others. However, the success of these tasks depends on the quality of the data used, since incorrect or incomplete data tend to impair the performance of the applied algorithms. Several academic data repositories do not contain or do not require the explicit information of the researchers' areas. In the data of the Lattes curricula, this information exists, but it is inserted manually by the researcher without any kind of validation (and potentially it is outdated, missing or even there is incorrect information). The present work utilized machine learning techniques in the inference of the researcher's areas based on the data registered in the Lattes platform. The titles of the scientific production were used as data source and they were enriched with semantically related information present in other bases, besides adopting other representations for the text of the titles and other academic information as orientations and research projects. The objective of this dissertation was to evaluate if the data enrichment improves the performance of the classification algorithms tested, as well as to analyze the contribution of factors such as social network metrics, the language of the titles and the hierarchical structure of the areas in the performance of the algorithms. The proposed technique can be applied to different academic data (not restricted to data present in the Lattes platform), but the data from this platform was used for the tests and validations of the proposed solution. As a result, it was identified that the technique used to perform the enrichment of the text did not improve the accuracy of the inference. However, social network metrics and numerical representations improved inference accuracy when compared to state-of-the-art techniques, as well as the use of the hierarchical structure of the classes, which returned the best results among the obtained.

Keywords: Text Classification. Text Enrichment. Research Interest. Topic Modeling. Lattes Platform.

Lista de figuras

Lista de algoritmos

Lista de quadros

Lista de tabelas

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Nas últimas décadas, a Web se tornou um enorme repositório de dados das mais diversas naturezas. Dentre esses dados, existem diferentes tipos de informações bibliométricas ou acadêmicas, tais como informações sobre universidades, currículos de pesquisadores e bibliotecas digitais.

Essas informações podem auxiliar na resolução de tarefas como a descoberta de especialistas em uma dada área, identificação de quais pesquisadores estão mais aptos a receber uma bolsa ou outro tipo de financiamento, sugestão de colaboradores para um projeto, entre outras diversas.

O sucesso destas tarefas depende da qualidade dos dados utilizados, pois dados incorretos, incompletos ou desatualizados tendem a prejudicar o desempenho dos algoritmos adotados (CANIBANO; BOZEMAN, 2009). Informações inseridas manualmente pelos usuários podem ser incompletas ou inconsistentes, pois, eles tendem a não preencher certas informações no perfil por não sentirem necessidade desse preenchimento ou mesmo por esquecimento ou desatenção (TANG et al., 2010).

Além disso, diversos repositórios de dados acadêmicos, e em especial bibliotecas digitais, não contêm informações referentes às áreas de atuação dos pesquisadores o que dificulta o teste e a validação de soluções automáticas para a inferência da potencial área de atuação de pesquisadores.

A Plataforma Lattes[1] pode ser considerada uma exceção a esta característica por possuir informações sobre a área de atuação (auto declarada pelos pesquisadores) de milhões de pesquisadores. Desde sua concepção, em meados dos anos 2000, a Plataforma Lattes, uma plataforma que gerencia currículos acadêmicos regida pelo órgão governamental CNPq, tem sido extremamente importante para a avaliação e reconhecimento de toda produção científica brasileira, sobretudo, por torná-la mais acessível à comunidade acadêmica. Além disso, uma vez que diversas avaliações de pesquisadores ou grupos de pesquisadores (como grupos de pesquisa e programas de pós-graduação) são avaliados considerando suas publicações, tal plataforma é suficientemente completa, no sentido que boa parte da produção científica de milhões de pesquisadores brasileiros se encontra registrada na mesma. Atualmente, a Plataforma Lattes contém informações de quase 5,5 milhões de

[1] ⟨http://lattes.cnpq.br/⟩

currículos, mais de 20 milhões de registros de artigos publicados e em cada currículo há na média 1,5 áreas de atuação declaradas.

Por gerenciar currículos com enfoque acadêmico, é possível a partir dos dados da Plataforma Lattes criar diferentes redes sociais acadêmicas, pois os currículos Lattes possuem informações de coautoria de artigos, orientações de diversas naturezas, áreas de atuação, entre outros, de forma que cada pesquisador pode ser visto como um nó da rede e as relações podem ser vistas como arestas na rede (por exemplo, relações de coautoria ou de orientação). Tais informações podem ser utilizadas, eventualmente, para resolver as tarefas citadas anteriormente, como por exemplo: o trabalho de Chagas, Perez-Alcazar e Digiampietri (2015) resolve o problema de reconhecimento de especialistas em uma dada área, Fonseca e Digiampietri (2016) fazem a obtenção de potenciais bolsistas produtividade e Digiampietri e Maruyama (2014) fazendo o reconhecimento de novas coautorias.

Para garantir a qualidade dos dados é necessária uma verificação da veracidade das informações inseridas em cada perfil, por exemplo, se um pesquisador declara ser coautor de outro, essa coautoria deveria ser verificada. Em particular, para as relações de coautoria, existe uma verificação dentro do próprio sistema, porém esta é limitada e o mesmo não ocorre com diversos outros campos.

Além disso, um pesquisador pode ter declarado atuar em uma dada área, mas essa informação não é necessariamente verdadeira. Isto pode influenciar no desempenho de técnicas de identificação de especialistas, caso a técnica leve essa informação em consideração.

Adicionalmente, conforme já mencionado, em várias bases bibliográficas nacionais e internacionais não estão disponíveis informações explícitas sobre as áreas de atuação dos autores. Assim, para se inferir automaticamente este tipo de informação, se faz necessário o desenvolvimento de uma solução assim como a proposta no presente trabalho.

Este trabalho propõe usar técnicas de aprendizado de máquina na inferência das áreas de atuação de pesquisadores, utilizando a produção científica como fonte de dados, bem como outras informações acadêmicas como orientações e projetos de pesquisa. Ademais, pretende-se enriquecer as informações dos currículos visando a obter uma estratégia que retorne resultados com melhor acurácia, testar o quanto representações numéricas para o texto impactam nos resultados, além de analisar a contribuição de fatores como métricas de redes sociais, idioma dos títulos e a própria estrutura hierárquica das áreas de atuação no desempenho dos algoritmos. Pretende-se validar o resultado das estratégias a serem

desenvolvidas com base nos dados dos bolsistas em produtividade, considerando que cada pesquisador atua na área do comitê que lhe concedeu a bolsa de produtividade em pesquisa.

Apesar de ser testada utilizando dados extraídos da Plataforma Lattes, a solução proposta nesta dissertação pode ser utilizada em outras bases de dados e visa também a detalhar a influência de diferentes características (como a tradução dos textos ou o enriquecimento) no processo de inferência de área de atuação.

O presente trabalho está organizado da seguinte forma: o restante deste capítulo discorre sobre os objetivos do mesmo. O capítulo 2 introduz conceitos básicos necessários para resolução da tarefa proposta. O capítulo 3 contém uma breve revisão bibliográfica da área. No capítulo 4, os materiais e métodos adotados nessa pesquisa são explicados. Os testes feitos são apresentados no capítulo 5 e seus resultados com suas devidas considerações são apresentados no capítulo 6. O capítulo 7 contém a conclusão do trabalho e possíveis trabalhos futuros.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é estudar e desenvolver diversas estratégias para a inferência das áreas de atuação de pesquisadores. As áreas de atuação declaradas nos Currículos Lattes dos bolsistas em produtividade em pesquisa serão usadas como *gold standard* no presente projeto.

1.1.2 Objetivos específicos

- Avaliar se o enriquecimento de dados utilizando informações presentes em base de dados exteriores ao Lattes (no caso a Wikipédia) melhora o desempenho dos algoritmos de classificação na tarefa de identificação das áreas de atuação.
- Estudar a contribuição que uma representação numérica do texto tem sobre o desempenho dos algoritmos de inteligência artificial.
- Analisar o quanto o uso de informações de hierarquia de classes pode contribuir para os resultados dos algoritmos.

2 Conceitos Fundamentais

Neste capítulo, conceitos fundamentais relacionados à área de mineração de texto e modelagem de tópicos serão apresentados, de forma a contextualizar algumas técnicas que serão adotadas nesse projeto. Para facilitar a compreensão, as próximas seções irão tratar de cada um dos tópicos abordados.

2.1 Mineração de Texto

Mineração de Texto, ou *Text Mining* em inglês, pode ser definida como uma forma de ajudar usuários a analisar e obter informações de grandes conjuntos de texto, usualmente encontrados na internet, para facilitar no processo de decisão. Seu objetivo principal é, então, descobrir padrões interessantes, incluindo tendências e *outliers*, em dados textuais (AGGARWAL; ZHAI, 2012).

Segundo Aggarwal e Zhai (2012), algumas características presentes em dados de texto afetam quais técnicas de mineração podem ou não ser usadas sobre eles, sendo a mais importantes destas, o fato que textos são esparsos e possuem alta dimensionalidade. Por exemplo, um *corpus* de texto pode ter um dicionário com 100000 palavras, porém um dado documento contém apenas algumas centenas.

Assim, o *corpus* pode ser representado como uma matriz termo-documento (*term-document matrix* em inglês) de tamanho n x d, na qual n é o número de documentos e d é o tamanho do dicionário. A entrada na posição (i, j) da matriz é a frequência normalizada da j-ésima palavra no documento i (AGGARWAL; ZHAI, 2012). O tamanho do dicionário tem grande influência nos algoritmos, pois, no exemplo dado anteriormente, 100000 palavras significaria o mesmo número de atributos na matriz, o que torna essa representação inviável para algumas abordagens, necessitando, por exemplo, da aplicação de técnicas de redução de dimensionalidade.

Além disso, dados de texto podem ser analisados em diferentes níveis de representação. Por exemplo, serem tratados como *bag-of-words* ou como uma *string* de palavras. Contudo, ambas não conseguem representar, de forma ideal, a semântica das palavras, o que facilitaria na mineração dos dados. Consideraremos nesse trabalho a representação

bag-of-words como a adotada, além do peso TF*IDF como forma de cálculo dos pesos de cada palavra nos documentos.

2.1.1 Term Frequency*Inverse Document Frequency (TF*IDF)

Segundo Aggarwal e Zhai (2012), uma das abordagens possíveis para cálculo do peso das palavras se baseia na probabilidade destas aparecerem nos documentos. Contudo, tal abordagem depende de uma lista de *stopwords* para eliminar palavras muito comuns e pouco relevantes. Decidir quais incluir nessa lista não é uma tarefa trivial, por isso atribuir pesos TF*IDF a elas pode ser uma melhor alternativa.

Os pesos TF*IDF fazem uso de um *corpus* de *background*, que é uma coleção de documentos normalmente do mesmo gênero que o documento que está sendo avaliado, como indicador do quão frequentemente uma palavra pode aparecer em um texto arbitrário (AGGARWAL; ZHAI, 2012).

A única informação adicional, além da frequência dos termos $c(w)$ que é necessária para computar o peso de uma palavra w que aparece $c(w)$ vezes nos textos de entrada, é computado o número de documentos ($d(w)$) em um *corpus* de background de D documentos, que contenham essa palavra (AGGARWAL; ZHAI, 2012). Desta forma é possível computar a frequência inversa dos documentos usando a equação 1. Em vários casos, $c(w)$ é dividido pelo máximo de ocorrência de qualquer palavra no documento, o que normaliza esse valor.

$$TF * IDF_w = c(w)log(\frac{D}{d(w)}) \tag{1}$$

O interessante dessa abordagem é que palavras que aparecem na maioria dos documentos terão o IDF perto de zero (AGGARWAL; ZHAI, 2012). O TF*IDF das palavras, então, são bons indicadores da importância delas, além de serem simples de computar e, assim, incorporados de uma forma ou de outra nos sistemas mais recentes (AGGARWAL; ZHAI, 2012).

2.2 Modelo de tópicos

A necessidade da análise de grandes conjuntos de texto fez com que a aplicação de modelos estatísticos hierárquicos baseados em tópicos se tornasse viável, por estes

permitirem que as tarefas fossem realizadas mais rapidamente. Formalmente, um tópico é uma distribuição de probabilidade sobre os termos do vocabulário estudado (BLEI; MCAULIFFE, 2008). Informalmente, pode-se dizer que um tópico representa um tema semântico do texto, no qual um documento que possui um grande número de palavras pode ser resumido a um número consideravelmente menor de tópicos (BLEI; MCAULIFFE, 2008).

Blei, Ng e Jordan (2003) argumenta que o objetivo, então, é encontrar pequenas descrições dos membros de uma coleção, permitindo o processamento eficiente de grandes coleções, enquanto preserva os relacionamentos estatísticos essenciais que são úteis para tarefas básicas, tais como, classificação, detecção de novidades (*novelty detection*), sumarização e cálculo de similaridade entre textos.

Contudo, não existe um consenso na literatura sobre qual seria a definição formal de modelo de tópicos (*Topic Model* em inglês) (NAVEED; SIZOV; STAAB, 2011). A definição informal de Naveed, Sizov e Staab (2011) considera modelo de tópico como um processo generativo, no qual documentos são criados e os padrões de ocorrência das palavras nos mesmos são armazenados em um *corpus* para produzir tópicos semanticamente coerentes.

Existe, atualmente, um grande número de abordagens usadas para fazer análise e modelagem de tópicos, porém, *Latent Dirichlet Allocation* (LDA) é considerada uma abordagem estado da arte por ser popular em análises de textos, devido a sua capacidade de produzir tópicos interpretáveis e semanticamente coerentes (NAVEED; SIZOV; STAAB, 2011). Assim, esta abordagem será utilizada neste trabalho.

2.2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) é um modelo probabilístico generativo de um *corpus*. Proposto por Blei, Ng e Jordan (2003), a ideia básica é que documentos são representados como misturas randômicas sobre tópicos e cada tópico é caracterizado por uma distribuição sobre palavras.

Algumas definições sobre o texto são usadas (BLEI; NG; JORDAN, 2003):

- Um documento é uma sequência de N palavras denotadas por $W = (W^1, W^2, ..., W^N)$.
- Um *corpus* é uma coleção de M documentos denotados por $D = \{D^1, D^2, ..., D^M\}$.
- Assume-se que a coleção possui Z tópicos ocultos $T = \{T^1, T^2, ..., T^Z\}$.

Desta forma, para cada palavra i no documento, adota-se $W^{(i)}$ para representar a i-ésima palavra na sequência, $D^{(i)}$ para representar o seu documento e $T^{(i)}$ o seu tópico. No LDA um documento $D^{(i)}$ é gerado por meio de uma distribuição de tópicos $T^{(i)}$ de uma distribuição Dirichlet $Dir(\alpha)$, que determina qual palavra é atribuída a qual tópico no documento. A atribuição de tópicos é feita a partir de uma amostragem de um tópico $T^{(i)}$ seguindo uma distribuição Multinomial $Multi(\theta)$. Esse processo é repetido até todos os K tópicos serem gerados para a coleção (VO; OCK, 2015).

Desta forma, o LDA assume o seguinte processo para cada documento $D^{(i)}$ no *corpus* D (BLEI; NG; JORDAN, 2003):

1. Escolha o número de palavras N seguindo uma distribuição de $Poisson(\psi)$.
2. Escolha a distribuição sobre os tópicos θ seguindo uma distribuição de Dirichlet $Dir(\alpha)$.
3. Para cada palavra $W^{(i)}$ em $D^{(i)}$:
    a) Escolha um tópico $T^{(i)}$ seguindo uma distribuição $Multi(\theta)$.
    b) Escolha uma palavra $W^{(i)}$ com probabilidade multinominal condicionada sobre o tópico $T^{(i)}$ $(P(W^{(i)}|T^{(i)}, \beta))$.

A probabilidade de se obter um novo documento d com n palavras é descrita na equação 2 (VO; OCK, 2015). Nesta, $P(\theta|\alpha)$ é encontrada na distribuição Dirichlet parametrizada por α e $P(W^{(i)}|T^{(i)}, \beta)$ é a probabilidade de $W^{(i)}$ sobre o tópico $T^{(i)}$ parametrizado por β. O parâmetro α pode ser visto como uma observação a priori do número de vezes que cada tópico é amostrado em um documento antes mesmo que qualquer palavra do mesmo tenha sido observada. O parâmetro β é hiper parâmetro que determina o número de vezes que palavras são amostradas antes mesmo de qualquer palavra do *corpus* ter sido observada (VO; OCK, 2015).

$$P(d|\alpha, \beta) = \int P(\theta|\alpha) \left(\prod_{i=1}^{n} \sum_{T^{(i)}} P(W^{(i)}|T^{(i)}, \beta) P(T^{(i)}|\theta) \right) d\theta \tag{2}$$

Uma vez que a distribuição é conhecida, pode-se encontrar a probabilidade para o *corpus* D como um todo, multiplicando as probabilidades de cada documento, como visto na equação 3.

$$P(D|\alpha, \beta) = \prod_{i=1}^{M} P(d_i) \tag{3}$$

Para inferir α e β a abordagem *Gibbs Sampling* (GRIFFITHS; STEYVERS, 2004) é adotada. *Gibbs Sampling* é um método para extrair tópicos de um *corpus* e encontrar a distribuição dos mesmos. A fórmula completa para o cálculo de $P(t_i|t_{-i}, w)$ é definida na equação 4 (GRIFFITHS; STEYVERS, 2004). Nesta equação, $n_{-i,j}^{(d_i)}$ representa o número de palavras do documento d_i atribuídas ao tópico j, não incluindo a atual (i) e $n^{(d_i)}$ o número total de palavras no documento d_i. Da mesma forma, $n_{-i,j}^{(w_i)}$ é o número de instâncias da palavra w atribuídas ao tópico j, não incluindo a atual (i) e $n_{-i,j}^{(.)}$ é o número total de palavras atribuídas ao tópico j, não incluindo a atual (GRIFFITHS; STEYVERS, 2004).

$$P(t_i|t_{-i}, w) = \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(.)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i,j}^{(d_i)} + K\alpha} \tag{4}$$

Uma simplificação desta equação é dada pelo algoritmo abaixo. Esta simplificação é útil para entender conceitualmente como o algoritmo funciona, porém ela não está completa, pois omite alguns cálculos presentes em Blei, Ng e Jordan (2003). No algoritmo, $P(t|d)$ representa a proporção de palavras no documento d pertencentes ao tópico t, enquanto que $P(w|t)$ define a proporção de tópicos t em todos os documentos que possuem a palavra w. Assim, a probabilidade a posteriori $P(t_i|t_{-i}, w)$ é dada por $P(t|d) * P(w|t)$. Ao observar o que a equação em si significa, é possível verificar que a mesma pode ser aproximada por $p(t|d) * p(w|t)$, o que valida a aproximação utilizada no algoritmo 1.

Algoritmo 1 *Gibbs Sampling*

```
1: procedure GIBBS SAMPLING
2:     Para cada documento d atribui-se randomicamente um dos K tópicos
   avaliados para cada palavra w
3:     do
4:         for cada documento d em D do
5:             for cada palavra w em d do
6:                 for cada tópico t do
7:                     Calcula-se P(t|d) e P(w|t)
8:                     Atribua w a um novo tópico t, na qual t tem
   probabilidade P(t|d) * P(w|t) de ser escolhido
9:     while a distribuição de tópicos não tiver convergido
```

3 Revisão Bibliográfica

A revisão bibliográfica deste trabalho objetivou fazer um levantamento de trabalhos correlatos ao proposto, no sentido de encontrar aqueles que lidam com a tarefa de inferência das áreas de atuação de pesquisadores usando abordagens diversas, de forma a levantar as principais contribuições dos mesmos. Considerou-se trabalhos correlatos, neste caso, publicações que lidam diretamente com a tarefa de modelagem dos interesses dos usuários, sejam estes interesses acadêmicos ou não. Nota-se que a proposta da pesquisa em si é inferir as áreas de atuação de pesquisadores, porém, como áreas de atuação podem ser compreendidas como os interesses de tais pesquisadores, trabalhos que fazem a modelagem destes interesses são correlatos, pois suas abordagens podem ser utilizadas para modelar as áreas de atuação.

Dentre as abordagens encontradas, existem duas vertentes que são proeminentes. A primeira usa técnicas não supervisionadas de aprendizado de máquina, normalmente variantes do *Latent Dirichlet Allocation (LDA)*, para descoberta de tópicos de interesses do usuário usando dados gerados pelo mesmo (*tweets*, produção científica, entre outros). A segunda abordagem lida com algoritmos supervisionados para fazer a classificação desses interesses.

A principal diferença entre ambas está na forma como os interesses são descobertos: a primeira não possui informação a priori sobre quais são os possíveis interesses, enquanto a segunda limita tais possibilidades para alguns interesses específicos. Destaca-se que a segunda abordagem também pode fazer uso do LDA, mas para a resolução de questões como redução de dimensionalidade ou enriquecimento de informação.

Os quadros 1 e 2 apresentam um resumo dos trabalhos descritos neste capítulo, separando os trabalhos que usam abordagens não supervisionadas para modelagem dos interesses dos usuários (quadro 1) daqueles que usam abordagens supervisionadas (quadro 2). Contudo, uma discussão mais detalhada sobre cada um deles será feita no decorrer deste capítulo.

Seguindo a linha não supervisionada, Naveed, Sizov e Staab (2011) apresentam uma extensão do LDA tradicional, denominada *Author-Topic-Time model* (ATT), que modela a evolução dos interesses do usuário com o passar do tempo. Tal proposta melhora o conteúdo dos documentos de texto com dados referentes ao autor e à data em que

Quadro 1 – Resumo dos trabalhos correlatos (abordagens não supervisionadas)

Artigo	Uso	Importância para o trabalho
Naveed, Sizov e Staab (2011)	Modificação do LDA que modela os interesses dos usuários com o passar do tempo	Adota uma base de dados acadêmica para testar a abordagem proposta
Xu et al. (2014)	Modificação do LDA que modela os interesses dos usuários com o passar do tempo	Adota uma base de dados acadêmica para testar a abordagem proposta
Katsurai, Ohmukai e Takeda (2016)	Representa os interesses dos usuários em tópicos obtidos por meio do resumo de sua produção científica, para realizar a desambiguação de nomes	Apresenta uma possível aplicação dos interesses do usuário
Xu et al. (2011)	Propõem uma modificação do LDA para modelar interesses de usuário por meio de seus tweets	Mostra como especializar uma extensão do LDA para uma plataforma específica

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

os documentos foram gerados, de forma que tanto autores quanto os *timestamps* são considerados como parte da entrada.

Para fazer a modelagem dos autores, cada componente (autor, texto e tempo) é modelado como uma distribuição sobre tópicos e cada tópico é modelado como uma distribuição sobre palavras. O modelo em si tem três conjuntos de parâmetros (diferentemente do LDA que possui apenas dois parâmetros): a distribuição dos autores sobre os tópicos (θ), a distribuição de tópicos sobre as palavras (ϕ) e a distribuição dos tópicos com o passar do tempo (ψ) (NAVEED; SIZOV; STAAB, 2011) . O artigo aplicou *Gibbs Sampling* para as inferência dos parâmetros citados .

Avaliou-se a eficácia do modelo em predizer autores e capturar o tempo de vida dos tópicos por meio da aplicação do mesmo em um subconjunto de resumos de publicações científicas do CiteSeer, cujos resultados foram comparados com o LDA padrão. Os dados usados do CiteSeer consistem nos resumos e títulos de artigos de 150 publicações entre os anos 2001 a 2009. 18 autores foram escolhidos aleatoriamente para a coleta de documentos. Um pré-processamento do texto foi feito para remoção de *stopwords* e *stemming* e separou-se o conjunto de dados em um conjunto de treinamento e um conjunto de teste. Para fazer a avaliação dos resultados, medidas como o *KL Divergence* entre o ATT e o LDA foram adotadas. Observou-se que o ATT retornou tópicos mais distintos entre si (NAVEED; SIZOV; STAAB, 2011).

Quadro 2 – Resumo dos trabalhos correlatos (abordagens supervisionadas)

Artigo	Uso	Importância para o trabalho
Vo e Ock (2015)	Propõem uma maneira de classificar artigos se baseando somente em seus títulos, empregando bases de dados exteriores para enriquecê-los por meio de tópicos	Propõe o enriquecimento de texto aplicado nessa dissertação
Chen e Li (2016)	Classificação de artigos usando os seus títulos enriquecidos a partir de uma base de dados externa com representações diversas para o texto	Outra forma de enriquecimento de texto para classificação de artigos, porém não considerando o contexto das palavras
Phan, Nguyen e Horiguchi (2008)	Abordagem baseada no LDA para fazer enriquecimento de texto	Enriquecimento de texto usando uma base de dados externa
Wang, Ma e Zhang (2016)	Usa o Word2Vec incorporado ao LDA para fazer o relacionamento entre os documentos e os tópicos	Forma de representação numérica para o texto que adota o LDA e Word2Vec como intermédio
Gabrilovich e Markovitch (2007)	Explora informações presentes na Wikipédia para conseguir fazer uma representação semanticamente relacionada do texto	Forma de representação numérica para o texto usando artigos do Wikipédia como intermédio
Miyata, Kano e Digiampietri (2013)	Realizaram a classificação automática das áreas de atuação dos pesquisadores cujos currículos estão presentes na Plataforma Lattes	Baseline para o trabalho atual, pois apresenta os resultados da literatura para classificação das áreas de atuação de pesquisadores da Plataforma Lattes

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Para entender os resultados obtidos, é necessária a definição de alguns conceitos. Segundo Naveed, Sizov e Staab (2011), autores com uma alta probabilidade de estar em um tópico são chamados de *topic pioners*, autores que frequentemente mudam seu tópico são chamados de *trend setters* e autores que seguem a tendência de outros autores são chamados de *mainstream*. O modelo *Author-Topic-Time* foi capaz de identificar tais perfis dentre os autores testados.

Por fim, um problema deste modelo é que o algoritmo usado para a inferência dos atributos não é escalável em grandes conjuntos de dados, sendo necessário, segundo os autores, a criação de uma extensão do *Gibbs Sampling* especificamente para a resolução do ATT (NAVEED; SIZOV; STAAB, 2011).

Seguindo a mesma ideia, o trabalho proposto por Xu et al. (2014), no qual *Author-Topic over Time* (AToT) é descrito, tem uma abordagem bastante similar a proposta por

Naveed, Sizov e Staab (2011). A diferença entre ambos está na arquitetura da estrutura adicional inserida ao LDA, de forma a tornar possível a modelagem da passagem de tempo na descoberta de tópicos.

Neste, tenta-se encontrar *pioneiros*, *mainstream* e *laggards* (autores que seguem tendências com atrasos), cujo o foco é modelar a dinâmica das mudanças dos interesses de usuário, não somente o tempo em si. A ideia para o modelo é muito parecida com a proposta por Naveed, Sizov e Staab (2011): autores são distribuições de tópicos, tópicos são distribuições de palavras e o tempo segue uma distribuição Beta. *Gibbs Sampling* foi usado para a inferência dos atributos.

Na avaliação do método proposto, 13 anos (1987 a 1999) de textos completos de artigos do *Neural Information Processing Systems* (NIPS) foram utilizados como base de dados. Mostrou-se a mudança dos tópicos dos autores com o passar dos anos por meio das cinco melhores palavras dentro de cada tópico e os cinco autores com maior probabilidade de pertencer a um dado tópico, além de um histograma que mostra a evolução dos tópicos com o passar dos anos. Além disso, para avaliar a qualidade do poder preditivo, a medida de perplexidade foi adotada e por meio da mesma constatou-se que o desempenho do AToT tem uma perplexidade menor que o *Author-Topic Model* (AT) (uma outra extensão do LDA que modela autores) quando o número de tópicos for maior que 10, o que indica que o modelo AtoT é melhor que o AT.

Estes dois trabalhos (XU et al., 2014; NAVEED; SIZOV; STAAB, 2011) são relevantes para esta pesquisa por modelarem interesses de usuários levando em consideração a passagem de tempo. Porém, diferentemente destas abordagens, este trabalho não objetiva fazer uma modificação do LDA para fazer tal modelagem, pois a passagem do tempo considerada será a mudança do resultado da classificação feita sobre as áreas de atuação na janela de tempo testada (últimos 3 anos, últimos 5 anos e toda produção científica), cabendo ao *Latent Dirichlet Allocation*, no caso, ser empregado somente nas fases de enriquecimento semântico e classificação.

Katsurai, Ohmukai e Takeda (2016) lidam com a representação dos interesses dos pesquisadores em um conjunto de tópicos de pesquisa encontrados em grandes conjuntos de dados acadêmicos. Tal trabalho adota LDA para calcular a distribuição de tópicos sobre as palavras de cada resumo da produção científica dos pesquisadores. Isso faz com que cada artigo tenha um conjunto de tópicos próprios, sendo possível converter as características

textuais de um pesquisador em vários vetores de tópicos e sumarizar os interesses do mesmo por meio do centróide (média) dos vetores.

Por não utilizar do *Author-Topic Model* na modelagem, a abordagem do artigo necessita, inicialmente, estimar a proporção de tópicos de um dado artigo e então usá-la para modelar os interesses do pesquisador. O cálculo dos centróides é interessante, segundo o trabalho, porque pode ser aplicado para calcular a relação entre dois pesquisadores r_1 e r_2 e assim explorar a similaridade entre os seus vetores de tópicos m_1 e m_2 (KATSURAI; OHMUKAI; TAKEDA, 2016).

Aplicou-se então esta similaridade na tarefa de desambiguação de nomes, que, no caso, lida com a identificação do autor correto de um artigo em caso de conflito de nomes. Tal desambiguação é feita do seguinte modo: computa-se o vetor de tópicos do artigo e sua distância ao vetor dos pesquisadores candidatos. O pesquisador com a menor distância para o vetor de tópicos do artigo é identificado como o seu autor (KATSURAI; OHMUKAI; TAKEDA, 2016).

Com o intuito de averiguar o desempenho da abordagem, alguns experimentos foram conduzidos utilizando o *CiNii Articles*, o maior conjunto de dados acadêmico japonês. Os resultados dos experimentos mostram que o método proposto obtém um melhor desempenho em desambiguação de nomes do que métodos que lidam diretamente com dados textuais (*Vector Space Model*) ou com meta dados (Afiliação, Nome, Similaridade do título do artigo com as publicações dos autores e palavras chaves), por alcançar uma acurácia de 92,60%.

Katsurai, Ohmukai e Takeda (2016) mostram como empregar a modelagem dos interesses de usuário (e subsequentemente das áreas de atuação) para fazer a desambiguação de nomes. Assim, esse trabalho é relevante para esta dissertação por apresentar uma possível aplicação da pesquisa proposta nesta dissertação, além de fornecer ideias de outros usos para os tópicos modelados pelo LDA (apesar destes não serem usados nessa pesquisa).

Interesses de usuário não necessariamente são apenas relacionados a produção científica e é importante reconhecer outras formas possíveis do uso de *topic model*. Xu et al. (2011) propõem uma modificação do *Author-Topic Model*, denominada *twitter-user model*, na qual modela-se os interesses por meio dos *tweets*, lidando com o problema destes possuírem muito ruído.

Author-topic model é uma extensão do LDA que inclui informações de autoria. Tal modelo assume que cada autor é representado como uma distribuição sobre tópicos e cada

palavra é associada a duas variáveis latentes: uma para um autor e outra para um tópico. Segundo Xu et al. (2011), inicialmente o modelo escolhe um autor de uma lista de autores possíveis e então sorteia um tópico relacionado ao autor escolhido para obter uma palavra usando a distribuição de palavras associadas a esse tópico.

Um dos principais desafios do trabalho é saber quais *tweets* são relacionados aos interesses dos autores, o que pode ser uma tarefa difícil, pois o *Twitter* é uma mistura de plataforma de informações com rede social e, assim, grande parte dos dados podem possuir ruído (XU et al., 2011). Mais especificamente, existem dois problemas no *Twitter*: diferentes intenções dos usuários quando usam a plataforma (uns a consideram como um instrumento de comunicação entre os amigos, enquanto outros a usam como um meio de disseminar informação) e o tamanho dos *tweets* (no máximo 140 caracteres, o que inviabiliza adotar muitas das estratégias baseadas em *bag of words*).

Os autores realizaram um estudo para avaliar o impacto que cada um dos problemas citados tem sobre a tarefa de identificação dos interesses dos usuários. *Retweet*, *reply*, *link* e *tag* são importantes, pois podem fornecer indicativos se um tweet é importante ou não (XU et al., 2011). Segundo um experimento feito, 81% dos *retweets*, 26% dos *replies*, 93% dos links e 58% das *hashtags* são considerados relacionados ao interesse dos usuários.

O comportamento do autor para fazer publicações também deve ser considerado. Por exemplo, um *tweet* sobre *iphone* provavelmente será relacionado aos interesses de um usuário que faz regularmente *tweets* sobre *smartphone*. Em contrapartida, se o usuário raramente faz esse tipo de postagem, isso pode implicar fatores externos como por exemplo um amigo comprou um *iphone* (XU et al., 2011).

A proposta desse trabalho tenta incorporar, então, os fatores citados anteriormente dentro de uma extensão do ATT, para capturar a real motivação destes *tweets*, ou seja, o interesse dos seus autores. O modelo proposto assume que cada *tweet* é associado a uma variável latente, que representa se ele está relacionado aos interesses do autor ou não e a origem do *tweet* é dada por ou uma distribuição de tópicos do autor ou por uma outra distribuição qualquer. A resolução dessas variáveis pode ser feita por meio do uso do *Gibbs Sampling*.

Com o intuito de avaliar o modelo proposto, um experimento utilizando dois meses de uso do *Twitter* de uma lista pré definida de usuários foi feito. Comparou-se o modelo proposto com ambos LDA e ATT usando duas medidas: perplexidade e as *top words* descobertas. Nos testes envolvendo a perplexidade dos modelos, cada um foi testado cinco

vezes com inicializações aleatórias e o valor resultante é a média da perplexidade de cada teste. Observou-se nos testes que o *Twitter-user model* tem perplexidade média menor que o ATT, que por sua vez tem uma perplexidade menor que o LDA. Como quanto menor a perplexidade, melhor é o modelo, isto sugere que o *Twitter-user Model* é melhor que os outros dois avaliados. Da mesma forma, os testes envolvendo avaliar as *top words* de cada tópico indicaram que o *Twitter-user model* tem um desempenho melhor que as outras técnicas. O método proposto por esse artigo pode ser considerado um trabalho prévio para várias outras tarefas a serem feitas no *Twitter*, por exemplo, recomendação de amigos, ranking de usuários e análise de redes sociais.

O *Twitter-user model* mostra como especializar uma extensão do LDA para uma plataforma, no caso o *Twitter*, de forma a melhorar o desempenho da modelagem de tópicos especificamente para a mesma. Apesar de fugir inicialmente da proposta de pesquisa desta dissertação, o fato de conseguir modelar interesses do usuário seguindo uma abordagem que utiliza uma extensão do LDA o torna um trabalho correlato ao proposto.

Paul e Girju (2009) fazem uma análise de áreas de pesquisa em linguística, classificando artigos se baseando no seu tópico. Esse trabalho utiliza uma abordagem semi-supervisionada que combina o algoritmo *Naive Bayes* com um classificador baseado na similaridade de um artigo com uma página da Wikipédia que descreve um dado tópico e adota o LDA para induzir tópicos de artigos que não possuíam classes (fase não-supervisionada). Nota-se que um dos objetivos desse trabalho foi apresentar como os tópicos evoluem com o passar do tempo em cada área de pesquisa estudada.

O conjunto de dados utilizado consiste de 4.700 artigos (1965 - 2008) da *ACL Anthology*, 1.700 artigos de revistas de Educação (1975 - 2008) e 2.300 artigos de revistas de linguística (1977 - 2008). A ideia na escolha dos artigos para o conjunto foi representar adequadamente cada área avaliada por meio das melhores revistas de linguística e educação e conferências que tenham uma grande abrangência em suas respectivas áreas (PAUL; GIRJU, 2009) . Parte das categorias referentes a cada artigo foram obtidas automaticamente das revistas, enquanto outra parte foi categorizada manualmente pelos autores, com o intuito de melhorar o conjunto de treinamento nas classes pouco representadas. No fim deste processo inicial, 86 classes distintas foram obtidas.

Inicialmente, a categorização dos artigos foi feita por meio do classificador *Naive Bayes*, que computa a probabilidade de um documento D ser da classe c_j ($P(D|c_j)$). O espaço de entrada para o algoritmo consiste das palavras presentes nos títulos e resumos,

bem como bigramas das mesmas. Permite-se que um artigo possa pertencer a múltiplas categorias ou nenhuma. Assim, um artigo é atribuído a um subconjunto de categorias que tenha um *z-score* acima de um certo limiar (PAUL; GIRJU, 2009).

Um *corpus* de conhecimento de *background* foi adotado, então, para corresponder exemplos sem categoria com aqueles que possuem uma, sendo este construído a partir de artigos da Wikipédia relacionados a linguística e educação. A medida cosseno é adotada para calcular a similaridade entre um artigo e uma página do Wikipedia e esta similaridade é usada para melhorar a probabilidade $P(D|c_j)$ do *Naive Bayes*. Os resultados desta abordagem tiveram um desempenho inferior ao *Naive Bayes*, porém, conseguiram acertar exemplos que o *Naive Bayes* não conseguiu. Além disso, uma vez que a maioria dos artigos no ACL Anthology não eram categorizados, foi necessário se utilizar do *Latent Dirichlet Allocation* para fazer a inferência dos seus tópicos.

Para computar as mudanças de um tópico com o passar do tempo, o *least square linear regression* foi aplicado sobre os pontos de cada tópico, para verificar o quanto cada um tem de tendência a altas ou baixas. Descobriu-se que em linguística os tópicos tendem a flutuar ano a ano mas não possuem uma tendência fixa, enquanto que na área de educação existe uma forte tendência de alta em tópicos sobre linguagem e leitura (PAUL; GIRJU, 2009).

O trabalho proposto por Paul e Girju (2009) é relevante para esta pesquisa, por fazer a classificação de quais áreas um determinado artigo se encontra (uma extensão disso pode ser adotada para mapear as áreas de atuação de um pesquisador, ao fazer um voto majoritário sobre as áreas de seus artigos). As principais diferenças com a presente pesquisa são: que esse trabalho lida com classificação semi supervisionada e não faz nenhum tipo de enriquecimento nos dados. Além disso, ele adota o LDA na fase não supervisionada para inferir a classe de dados que não possuem uma. Nesta dissertação de mestrado, utiliza-se de uma ideia similar para auxiliar no processo de classificação, seja por meio do uso de tópicos oriundos do LDA para representar os dados ou para enriquecer os dados com informações semanticamente relacionadas.

Vo e Ock (2015) propõem uma maneira de classificar artigos se baseando somente em seus títulos, empregando bases de dados exteriores (no caso DBLP, LNCS e Wikipédia) para enriquecê-los e usando palavras recorrentes em tópicos descobertos dessas bases de dados por meio do LDA.

Segundo Vo e Ock (2015), classificar documentos de texto curtos é desafiador devido a sua esparsidade e por serem limitados em ambos ocorrência das palavras e contexto das mesmas. Visto isso, títulos de produções científicas são considerados pequenos fragmentos de texto, portanto sendo necessário algum tipo de abordagem para melhorar o desempenho dos algoritmos de classificação, sendo a escolhida por esse artigo fazer um enriquecimento semântico usando uma base de dados externa.

O artigo adota duas formas de se fazer o enriquecimento: a primeira usando tópicos oriundos da modelagem de tópico das bases de dados externas como atributos externos e a segunda combinando palavras de tópicos adaptados (nome adotado pelo artigo para chamar tópicos atribuídos aos títulos) as do texto original para enriquecê-lo. Ambas abordagens usufruem de tópicos obtidos por meio do uso de *Latent Dirichlet Allocation* utilizando dados oriundos das bases DBLP (2 milhões de artigos de Ciência da Computação obtidos em dezembro 2012), LNCS (resumos de 43.600 artigos) e Wikipédia (42.000 artigos).

Os tópicos foram estimados usando o GibbsLDA++[1]. Os valores de α, β foram respectivamente $\alpha = 50/T$ e $\beta = 0.01$, sendo T o número de tópicos avaliados. Os números de tópicos testados foram: 20, 30, 40, 50, 60, 70, 80, 90, 100, 120 e 150.

Seja $\overline{t} = t_1, t_2, ..., t_i$ e $\overline{d} = w_1, w_2, ..., w_i$ respectivamente o vetor de tópicos gerados pelo LDA e o vetor de documentos de texto, os tópicos dos documentos de texto são combinados por meio de um *matching* entre as palavras nos documentos de texto e as palavras do modelo de tópicos. Para enriquecer o texto, usa-se palavras dos tópicos adaptados dentro do texto original. Contudo, é necessário avaliar quais são os melhores tópicos para cada palavra, dado o contexto do título. Para tal, para cada palavra em t_i, adiciona-se a probabilidade dos tópicos atribuídos a essa palavra a probabilidade de tais tópicos a outras palavras no mesmo documento, de forma a seguir a equação 5. Determina-se os melhores tópicos adaptados por meio de um ranking $Rank_n(P_w(t_i))$

$$\overline{P_w(t_i)} = P_w(t_i) + \sum \beta P(t_i) \tag{5}$$

Na equação 5, $\overline{P_w(t_i)}$ é a probabilidade de um tópico adaptado t_i ser diretamente atribuído a palavra w e $\sum \beta P(t_i)$ é a soma das probabilidades deste tópico t_i das outras palavras presentes no documento. O atributo β controla o impacto dos outros tópicos adaptados sobre a escolha do que será enriquecido. Após selecionar os melhores tópicos

[1] ⟨http://gibbslda.souceforge.net⟩

usando o ranking das melhores probabilidades calculadas usando a equação 5, combina-se k palavras dos melhores n tópicos ao texto original do título.

Os classificadores *Support Vector Machine* (SVM), *Naive Bayes* (NB) e *K-nearest Neighbor*(KNN) foram adotados para provar que a abordagem descrita é superior a somente utilizar títulos sem enriquecimento na classificação. O TF-IDF foi escolhido para a representação do texto dos títulos e os seus valores foram os utilizados como entrada para os algoritmos citados anteriormente.

A avaliação utilizou de títulos coletados de revistas e anais de eventos científicos relacionados à área de Ciência da Computação, resultando em 1.400 documentos. Seis foram as classes escolhidas para a avaliação: Bioinformática, Arquitetura de Computadores, Banco de Dados, Sistemas de Informações Geográficos, Redes e Processamento de Linguagem Natural. Acurácia, Precisão, Medida-F e Revocação foram adotadas para computar a eficácia da proposta. No geral, o algoritmo SVM obteve os melhores resultados.

Os resultados vistos em Vo e Ock (2015) indicam que o enriquecimento melhorou consideravelmente o desempenho dos algoritmos. Por exemplo, a acurácia do *baseline* usando SVM era de 71,61%, enquanto que o enriquecimento usando a Wikipédia, por meio dos tópicos como atributos externos, resultou em uma acurácia de 76,56% e o enriquecimento usando palavras externas dos tópicos resultou em uma acurácia de 76,76%.

Segundo Vo e Ock (2015), a sua abordagem torna as características dos documentos de texto curtos mais concisos devido às palavras externas, pois estas afetam a correlação entre o relacionamento das palavras, o que melhora a classificação. Além disso, aplicar o LDA faz o texto menos esparso e mais orientado a um determinado tópico. Nota-se que este trabalho é o alicerce para o proposto nessa dissertação, pois, partiu-se de métodos semelhantes, mudando-se, nesse caso, a base de dados original (Lattes), quais dados exteriores serão usados (Wikipédia) e alguns dos algoritmos que serão testados na classificação. A diferença primária entre ambos, se dá no fato que Vo e Ock (2015) faz a classificação de artigos, enquanto que esta pesquisa classifica a área de atuação dos pesquisadores como um todo, sendo portanto uma expansão da ideia anterior. Adicionalmente, a pesquisa da presente dissertação também utilizou características da análise de redes sociais.

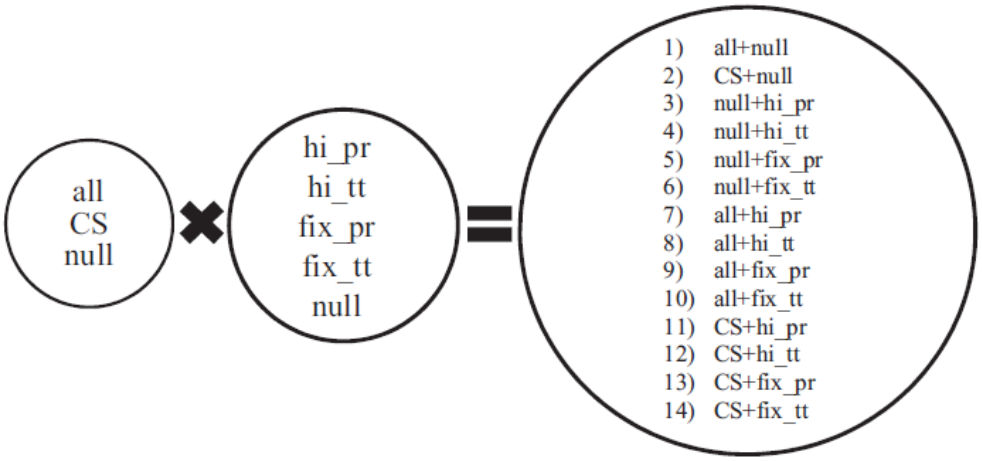
O trabalho de Vo e Ock (2015) não é o único a propor o enriquecimento do texto por meio do LDA, existindo diversos outros, como por exemplo o proposto por Chen e Li (2016), cuja diferença com Vo e Ock (2015) está no fato dele adotar diferentes formatos para as frequências dos termos e informação dos tópicos, combinado-as e testado-as empiricamente

em um conjunto de dados chinês e dois em inglês, de forma a avaliar qual seria a melhor abordagem. O conjunto de dados em chinês é o *Academia Sinica Balanced Corpus of Modern Chinese* (BCMC-V4), enquanto que os conjuntos em inglês são 20 *newsgroups* e RCV1-V2.

Duas formas de representar os documentos foram estudadas: características que usam frequência dos termos e características que usam informações dos tópicos. As características envolvendo a frequência dos termos consideram dois conjuntos diferentes de termos: todos do vocabulário (all) e aqueles usando a medida chi-quadrado (CS) para avaliar o quão informativa uma característica é.

Por outro lado, as representações envolvendo informações de tópicos foram geradas por meio do modelo LDA, usando uma abordagem hierárquica com 2, 4, 8, 16 e 32 tópicos respectivamente, possuindo então dois tipos de modelo: número fixo de tópicos (fix) e com uma estrutura hierárquica (hi). Além disso, existem dois tipos de formatos que foram adotados: a probabilidade do documento em cada tópico (pr) e os termos dentro dos tópicos (tt). A combinação de todas as características foi avaliada, considerando que existem dois tipos de frequência de termos, dois tipos de modelos de tópicos e duas formas de se codificar os tópicos, como visto na figura 1. Assim, foram criados 14 diferentes conjuntos de característica.

Figura 1 – Abordagem proposta por Chen e Li (2016)



Fonte: Chen e Li (2016)

Os 14 conjuntos de características diferentes foram testando usando *Support Vector Machine* e se constatou que informações derivadas de tópicos oriundos do LDA deveriam

ser integradas as informações de frequência de termo, uma vez que os testes envolvendo essas combinações tiveram o melhor desempenho de classificação.

Apesar de Chen e Li (2016) proporem abordagens interessantes sobre enriquecimento de texto, sobretudo a que usa um nível hierárquico de número de tópicos para o modelo LDA, a forma em que o texto é enriquecido difere de Vo e Ock (2015) ao combinar linearmente as características que usam frequência dos termos e características que usam informações dos tópicos e não considerar, por exemplo, o contexto de cada palavra do texto. Assim, o presente dissertação segue a abordagem de Vo e Ock (2015) ao invés de utilizar de Chen e Li (2016), por este não considerar o contexto das palavras.

Phan, Nguyen e Horiguchi (2008) mostram outro exemplo de artigo que utiliza do modelo LDA como forma de enriquecer o texto, lidando, no caso, com fragmentos pequenos de texto oriundos da internet, que usualmente são difíceis de classificar devido à esparsidade do texto, que não fornece informação suficiente para computar a ocorrência das palavras. Desta forma, a abordagem proposta neste utiliza de uma fonte de dados externa (chamada de conjunto de dados universal) para computar tópicos usando o modelo LDA, que então são empregados para enriquecer o texto o original.

A abordagem proposta no artigo é adotada em dois problemas distintos, sendo o primeiro a desambiguação de pesquisas na Web e o segundo a classificação de doenças usando fragmentos de textos médicos. Wikipédia e Medline foram os conjuntos adotados para resolver ambas as tarefas, respectivamente.

Para a construção do conjunto de dados universal, palavras-chave foram preparadas para serem pesquisadas na Wikipédia e, para cada uma, a página da web correspondente foi obtida bem como páginas relevantes citadas dentro dela (no máximo 4). Desta forma, o conjunto obtido possuía mais de 70.000 artigos.

A implementação GibbsLda++ (PHAN; NGUYE, 2007) foi adotada em Phan, Nguyen e Horiguchi (2008) para fazer a inferência dos tópicos, cujo número variou nos seguintes valores: 10, 20, 30, 100, 150 e 200. O algoritmo escolhido para fazer a classificação de texto foi *MaxEnt*, pois, segundo Phan, Nguyen e Horiguchi (2008), ele é robusto e tem sido aplicado em diversas tarefas de processamento de língua natural, além de usar o texto em si como entrada, ao invés de dados numéricos.

A integração dos tópicos obtidos com os dados originais em Phan, Nguyen e Horiguchi (2008) é dada da seguinte forma: dado um conjunto de documentos $\overline{W} = {w_n}^{M_{m=1}}$ e W os documentos no conjunto de dados universal. Por exemplo, $\overline{W}$ seriam os fragmentos

de texto de *queries* e W os documentos da Wikipédia. A inferência de tópicos é feita a partir de ambos W e $\overline{W}$ usando *Gibbs Sampling*, a diferença é que $\overline{W}$ precisa de um número menor de iterações para convergir (PHAN; NGUYEN; HORIGUCHI, 2008).

Sejam $\vec{w}$ e $\vec{z}$ os vetores de todas as palavras e a atribuição de tópicos para o conjunto W, $\vec{\overline{w}}$ e $\vec{\overline{w}}$ os vetores de todas as palavras e atribuição de tópicos para o conjunto $\overline{W}$. A atribuição dos tópicos para uma palavra t em $\overline{w}$ depende dos tópicos atuais para outras palavras em $\overline{w}$ e de todas as palavras em w. Entre outras palavras, a inferência das palavras nos documentos em $\overline{W}$ depende também dos tópicos encontrados para W.

Após fazer o *sampling* dos tópicos usando *Gibbs Sampling* e a inferência dos tópicos propriamente ditos, integram-se os tópicos obtidos e o documento original, fazendo uma discretização de cada tópico, nomeando-o de acordo com a probabilidade dele no texto. Um exemplo de *snipplet* enriquecido pode ser visto na Figura 2.

Figura 2 – Exemplo de enriquecimento da abordagem proposta por Phan, Nguyen e Horiguchi (2008)

(snipplet 1) online poker tilt poker money payment processing deposit money tilt poker account visa mastercard credit card atm check debit card topic:70 topic:103 topic:103 topic:103 topic:103 topic:137 topic:137 topic:188

Fonte: Phan, Nguyen e Horiguchi (2008)

Os resultados nas duas tarefas testadas foram satisfatórios, pois a acurácia aumentou de $79,84\%$ (sem o enriquecimento) para $83,73\%$ no caso da tarefa de desambiguação de pesquisas na Web e teve resultados similares para a tarefa de classificação de doenças usando fragmentos de textos médicos, com acurácia de $65,23\%$ com enriquecimento contra 65.68% sem ele. Nota-se que a abordagem com enriquecimento usou apenas 4.500 exemplos de treinamento, muito menor do que os 22.500 usados na abordagem sem enriquecimento (PHAN; NGUYEN; HORIGUCHI, 2008).

A diferença do trabalho de Phan, Nguyen e Horiguchi (2008), para o adotado nessa pesquisa (VO; OCK, 2015), se dá na forma como o enriquecimento é feito, na qual restringe o uso da abordagem, pois esta não funciona em algoritmos com entrada numérica (como SVM), por exemplo, necessitando de outra forma de integrar os dados (PHAN; NGUYEN; HORIGUCHI, 2008).

Wang, Ma e Zhang (2016) seguem uma abordagem análoga a representação numérica utilizada neste mestrado, porém incorporando o Word2Vec ao LDA para obter o relacionamento entre os documentos e os tópicos. Adicionalmente, as relações contextuais do texto

descobertas por meio do Word2Vec permitem o uso da distância euclidiana para calcular a distância entre os documentos e os tópicos.

Support Vector Machine (SVM) foi adotado como forma de análise da representação numérica encontrada no artigo, aplicando-o no conjunto de dados 20Newsgroups. Os resultados foram comparados com outras abordagens similares, tais como TF-IDF + SVM, Word2Vec + SVM e LDA+SVM.

No geral, descobriu-se que o TF-IDF é a única abordagem melhor do que a proposta por Wang, Ma e Zhang (2016), sendo 2% mais acurada, contudo, demorando quatro horas a mais para terminar a execução. Assim, a solução proposta pelos autores possui um desempenho eficiente entre os métodos considerados. Além disso, ao avaliar diferentes números de tópicos para o LDA, descobriu-se que o método proposto por Wang, Ma e Zhang (2016) requer uma quantidade maior de documentos do que o LDA puro, uma vez que a representação dos documentos leva mais informação contextual em consideração (WANG; MA; ZHANG, 2016). O artigo de Wang, Ma e Zhang (2016) é importante para este mestrado, pois é um exemplo de como fazer uma representação numérica do texto, similar a adotada em parte dos testes deste trabalho, sendo que esta será melhor explicada nas próximas seções.

Contudo, a representação de texto não se restringe somente ao uso do LDA, pois existem abordagens como o *Explicit Semantic Analysis* (GABRILOVICH; MARKOVITCH, 2007), por exemplo, que explora informações presentes na Wikipédia para conseguir fazer uma representação semanticamente relacionada do texto. Ou seja, é possível fazer uma representação de qualquer texto em termos de conceitos presentes na Wikipédia.

A justificativa para esta ideia é representar relações semânticas em termos de conceitos naturais, que são definidos por humanos e são facilmente explicados, o que difere dos conceitos latentes propostos presentes no *Latent Dirichlet Allocation*, que não são baseados em conceitos da psique humana, mas sim em relações probabilísticas dentro do texto (GABRILOVICH; MARKOVITCH, 2007).

Cada conceito da Wikipédia é representado como um vetor de palavras que ocorrem no artigo relacionado. As entradas desse vetor são atribuídas usando pesos TF-IDF e para tornar o processo mais rápido, constrói-se um índice invertido que mapeia cada palavra em uma lista de conceitos que ela aparece (GABRILOVICH; MARKOVITCH, 2007).

No geral, o algoritmo do ESA utiliza os seguintes passos: dado um fragmento de texto, ele é representado como um vetor TF-IDF. O interpretador semântico (que é

baseado em um classificador de centróides) itera sobre as palavras, obtendo as entradas correspondentes no índice invertido e as combina em um vetor ponderado de conceitos, que representa o dado texto. Seja $T = w_i$ o texto de entrada e $< v_i >$ o seu vetor TF-IDF, no qual v_i é o peso da palavra w_i. Seja $< k_j >$ a entrada no vetor invertido para a palavra w_i, no qual k_j quantifica a força da associação da palavra w_i com o conceito da Wikipédia c_j (c_j pertence a c_1,...,c_n), sendo N o número total de conceitos da Wikipédia. Então, o vetor semântico V para o texto T é um vetor de tamanho N, que computa o peso de cada conceito c_j como $\sum_{w_i \in T} v_i * k_i$. As entradas nesse vetor refletem a relevância dos conceitos correspondentes ao texto T. Para computar a relação semântica de um par de textos, computa-se a medida cosseno entre os seus vetores (GABRILOVICH; MARKOVITCH, 2007).

Exemplos de uso do ESA são os trabalho de Gabrilovich e Markovitch (2006) e Li et al. (2013), sendo que ambos apresentam que o ESA pode ser utilizado diretamente na tarefa de classificação, ou seja, cada conceito da Wikipédia pode servir como classe.

Uma dos problemas no ESA, observando os trabalhos citados anteriormente, é a necessidade de se utilizar do Wikipédia quase que em sua totalidade para que a abordagem funcione (tanto Gabrilovich e Markovitch (2006) e Li et al. (2013) usam uma grande parcela do Wikipédia), o que faz com que aplicar essa abordagem em algumas aplicações se torne inviável, por isso este não é adotado no presente trabalho.

Por fim, Miyata, Kano e Digiampietri (2013) realizaram a classificação automática das áreas de atuação dos pesquisadores cujos currículos estão presentes na Plataforma Lattes, usando uma abordagem de TF-IDF. O objetivo desse trabalho é fazer a combinação de técnicas de Mineração de Textos com Análise de Redes Sociais, na tentativa de identificar as áreas de atuação de pesquisadores usando os títulos de suas publicações e informações de suas redes de coautoria (MIYATA; KANO; DIGIAMPIETRI, 2013).

Para tal, utiliza-se da lista de bolsistas produtividades do ano de 2010 em todas as áreas de conhecimento, somente incluindo aqueles que declararam uma única "Área de Atuação" nos três níveis da taxonomia do Lattes (Grandes Áreas, Áreas e Subáreas), o que resultou em três conjuntos de dados:

- O referente às Grandes Áreas, com 9748 pesquisadores divididos em 8 classes.
- O referente às Áreas, com 7297 pesquisadores divididos em 76 classes.
- O referente às Subáreas, com 3427 pesquisadores com 443 classes.

Três características foram extraídas para inferir as áreas de atuação: uma baseada na mineração de texto e duas baseadas na rede de coautoria. A característica baseada na mineração de texto adota um valor calculado sobre o TF-IDF de cada classe, abordagem essa que será melhor explicada durante o estudo de caso. As duas características de Análise de Redes Sociais foram a porcentagem dos vizinhos pertencentes a cada "Área de Atuação" (seja grande área, área ou subárea) utilizando o primeiro e o segundo nível de vizinhança. Para ambas características, foram obtidos vetores para cada "Área de Atuação", nas quais o valor considerado é a porcentagem dos vizinhos que pertencem a mesma. Adotou-se o *holdout* para os testes, separando 90% do conjunto para treinamento dos classificadores e 10% para o teste. Os experimentos variaram o período dos dados de 1 a 10 anos e analisaram as taxas de acerto de cada um.

Os testes usando as Grandes Áreas como classes e usando somente Mineração de Texto, segundo o artigo, apresentaram que os resultados melhoram progressivamente ano após ano. O melhor resultado obtido adotou 9 anos de publicações e atingiu 86,67% de acerto. Integrar tais resultados à Análise de Redes Sociais fez a taxa de acerto subir para 90,56% (em uma combinação linear de Mineração de Texto com a análise de vizinhança de nível 2), e 90,87% usando o classificador *Rotation Forest* para 10 anos de publicações.

A mesma abordagem foi feita usando as Áreas como classe, na qual usar somente Mineração de Texto não obteve bons resultados. Nesse caso, adotar a combinação de Mineração de Texto com a análise de vizinhança de nível 2 trouxe o melhor desempenho (usando os 10 anos de publicação), resultando em 84,11% de acurácia. *Rotation Forest* resultou em 70% de acerto para o período de 10 anos.

Igualmente, os mesmos testes foram feitos usando as Subáreas como classe, nas quais adotar somente Mineração de Texto resultaram em resultados consideravelmente ruins (o melhor deles com 26,53% de acerto). Novamente, o melhor resultado encontrado foi aquele que adotou a combinação de Mineração de Texto com a análise de vizinhança de nível 2, resultando em uma taxa de acerto de 59,77%. O uso de *Rotation Forest* para combinar as medidas, nesse caso, resultou em uma taxa máxima de acerto de 36,15% para o período de 10 anos. Tais resultados apresentaram que a combinação simples da técnica de mineração de texto com análise de vizinha de nível 2 obteve melhores resultados do que aqueles produtivos pelo uso individual de cada técnica.

O presente trabalho pode ser considerado uma extensão deste artigo, pois espera-se ser possível melhorar os resultados obtidos com TF-IDF, por meio do enriquecimento dos

títulos e uso de outras técnicas, bem como o uso de informações de redes sociais, que podem ser extraídas da estrutura do Lattes.

4 Materiais e métodos

Este capítulo visa a detalhar os materiais e métodos desta dissertação, minuciando as etapas necessárias para a realização da inferência das áreas de atuação de pesquisadores presentes na Plataforma Lattes. Adota-se as áreas de atuação, no caso, como aquelas presentes nos três primeiros níveis da taxonomia de áreas do Currículo Lattes, de forma a avaliar os resultados em cada um dos níveis. Assim, os modelos consideram como classe as Grandes Áreas, Áreas e Subáreas.

A base de dados utilizada é a mesma adotada por Fonseca e Digiampietri (2016), na qual apenas pessoas contempladas com bolsas de produtividade em pesquisa do CNPq são consideradas. Parte-se da hipótese de que bolsistas produtividade declaram sua áreas de atuação corretamente e esta amostra foi utilizada, pois eles foram avaliados como "produtivos" por um comitê de área do CNPq. Desta forma, as áreas declaradas por eles foram utilizadas como referência para o treinamento de classificadores, por serem consideradas como *ground truth* para esta pesquisa. Os textos utilizados do currículo de cada pesquisador englobam tanto os títulos dos artigos publicados (seja em conferências ou periódicos), quanto os títulos dos projetos de pesquisa e das orientações.

Várias abordagens de treinamento e representação de texto foram avaliadas neste projeto com o intuito de identificar quais teriam o melhor desempenho sobre a tarefa de inferência das áreas de atuação dos pesquisadores. Para representação de texto se adotou o *bag of words* e a representação TF-IDF, de forma a utilizar dos valores obtidos no TF-IDF para classificar o texto.

Uma vez que o texto dos títulos da produção científica dos pesquisadores são fragmentos pequenos de texto, e se sabe que por causa do tamanho pequeno, este tipo de texto não fornece todas as informações necessárias de ocorrência ou contexto para terem boas medidas de similaridade (PHAN; NGUYEN; HORIGUCHI, 2008), a abordagem de enriquecimento de texto proposta por Vo e Ock (2015) foi adotada com o intuito de avaliar o desempenho dos algoritmos.

Ademais, representações numéricas baseadas no TF-IDF e no enriquecimento do texto também foram propostas. Assim, dois tipos de representação são criadas: uma baseada no valor TF-IDF dos títulos sobre cada uma das classes (algo análogo a classificação que Miyata, Kano e Digiampietri (2013) seguem), no qual ao invés de usar o TF-IDF para

classificar os títulos os utiliza para representar o texto; a outra que é baseada em tópicos oriundos de uma base de dados externa, que faz o mapeamento do texto para esses tópicos através de uma função.

Para as representações numéricas, os algoritmos *Naive Bayes*, SVM (com kernel linear), Árvore de Decisão e *Random Forest* foram utilizados como classificadores para a inferência das áreas de atuação. Foram utilizadas as implementações de Naive Bayes, Árvore de Decisão e Random Forest disponíveis no arcabouço Weka[1] e de SVM disponível na biblioteca LibSVM.

Os parâmetros de entrada para os algoritmos avaliados usaram os valores padrões presentes no arcabouço Weka 3.8, descritos a seguir. Para todos os conjuntos foi usado um *batch size* de 100, a Árvore de Decisão segue a implementação J48 (uma implementação da árvore C4) e usa como entrada um fator de confiança de 0,25, com um número mínimo de nós por níveis de 2. A *Random Forest* usou 100 iterações para treinamento e, por fim, a SVM adotou *kernel* linear com valores de entrada normalizados e a variável de custo $C = 1,0$.

Para efeito de avaliação do desempenho, testou-se os algoritmos de aprendizado de máquina citados usando a abordagem 10-*fold cross validation*, adotando os conjunto de dados com e sem enriquecimento dos títulos, e títulos traduzidos em cada um dos problemas (Grandes Áreas, Áreas e Subáreas).

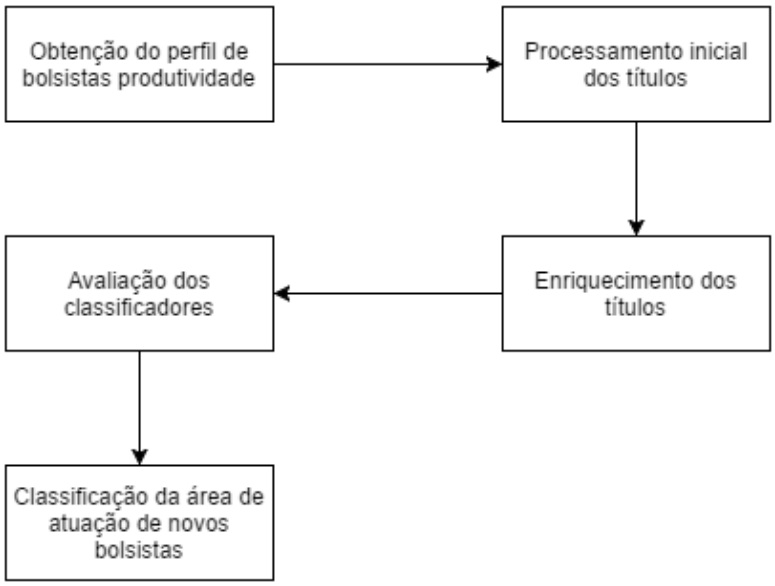
Avaliou-se, também, o quanto o período considerado nas análises contribui para o processo de classificação, além de também avaliar a contribuição da rede de relacionamentos (no caso, a rede de coautorias) para a inferência das áreas de atuação. Este processo parte do pressuposto que se um pesquisador colabora com diversos pesquisadores de uma mesma área, provavelmente ele também será dessa área.

A figura 3 sumariza arquitetura do sistema proposto, bem como as fases necessárias para sua execução. Nota-se que as etapas de "obtenção do perfil de bolsistas produtividade" e "processamento inicial dos dados" foram realizados anteriormente ao desenvolvimento desta dissertação, no conjunto de dados utilizados por Fonseca e Digiampietri (2016), cabendo a este trabalho apenas executar as fases subsequentes.

As próximas seções detalham cada uma das abordagens citadas, explicitando os métodos utilizados.

[1] Arcabouço WEKA: ⟨http://www.cs.waikato.ac.nz/ml/weka/⟩, acessado em 30/01/2018

Figura 3 – Estrutura proposta para inferência das áreas de atuação



Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

4.1 Base de dados

A base de dados do projeto utiliza dos dados de bolsistas produtividade, sendo esta similar a adotada por Fonseca e Digiampietri (2016), no sentido de utilizar do mesmo *snapshot* da plataforma Lattes. Três problemas distintos foram avaliados pelo trabalho, seguindo os três níveis da taxonomia do Lattes, nomeadamente a inferência das: Grandes Áreas, Áreas e Subáreas.

Desta forma, três conjuntos de dados foram construídos para cada um dos problemas avaliados. Destaca-se que em todos os conjuntos construídos apenas foram considerados pesquisadores que só declararam um valor em seu currículo na Plataforma Lattes para a classe avaliada, ou seja, só possuem um valor para as Grandes Áreas no conjunto das Grandes Áreas, um valor para as Áreas no conjunto referente as Áreas e um valor para as Subáreas no conjunto referente as mesmas. Nesse contexto, cada um deles possuí um número de classes distintos, sendo oito classes para as Grandes Áreas, 75 classes para as Áreas e 484 classes para as Subáreas. A relação do número de pesquisadores presentes em cada um dos conjuntos pode ser vista na tabela 1.

Os títulos das publicações científicas, projetos e orientações de cada um dos pesquisadores foram representados por meio de um *bag of words*, tendo remoção de *stopwords* aplicadas sobre o texto por meio de uma biblioteca da linguagem Python chamada *nltk*. O texto resultante possuí três formas diferentes para sua representação: original, traduzido e

Tabela 1 – Número de pesquisadores por conjunto de dados

Conjunto	Número de Pesquisadores
Grandes Áreas	9351
Áreas	6792
Subáreas	4060

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

traduzido/enriquecido. Os títulos com texto original possuem o mesmo texto presente na Plataforma Lattes, com a diferença de ter *stopwords* removidas dos mesmos.

Títulos com texto traduzidos, como o próprio nome diz, são aqueles cujo texto foi traduzido para português por meio da biblioteca *TextBlob* do Python. A motivação por trás da tradução se dá pelo fato de que o Lattes é uma base de dados com texto misto, ou seja, existem títulos escritos em línguas diversas (português, espanhol, inglês e outros). Isso faz com que o número de características (tamanho do vocabulário) seja consideravelmente grande, podendo atrapalhar no desempenho dos algoritmos de mineração de dados. Assim, realizar a tradução prévia dos títulos reduz o tamanho do vocabulário e pode ser vantajoso para os algoritmos.

Já os títulos trazidos e enriquecidos são aqueles que tiveram seu texto traduzido previamente (conforme apresentado no parágrafo anterior) e, então, foi utilizada uma base de dados externa para seu enriquecimento. Foram utilizados textos em português da base externa, e assim, para o enriquecimento funcionar adequadamente foi necessário que o texto dos títulos estivesse em português também (conforme será detalhado adiante).

4.2 Base de dados externa

A base de dados externa aplicada no projeto se baseou em páginas da Wikipédia para sua construção, seguindo uma abordagem similar à utilizada por Phan, Nguyen e Horiguchi (2008). As classes adotadas pelo projeto nos três problemas (Grandes Áreas, Áreas e SubÁreas) foram usadas como palavras-chaves para buscar páginas relacionadas a elas na Wikipédia. Desta forma, para cada palavra-chave, as 10 primeiras páginas retornadas na busca foram obtidas de maneira automática utilizando a biblioteca *wikipedia* do Python. Isto resultou em 28.741 artigos da Wikipédia. Nota-se que nem todos foram utilizados porque algumas buscas resultaram em artigos iguais, assim artigos duplicados foram desconsiderados resultando em 15.047 artigos distintos. O texto das páginas também

passou por um processo de remoção de *stopwords* assim como os títulos da produção científica no Lattes.

O *corpus* obtido sobre o texto tratado foi usado como entrada no algoritmo *Latent Dirichlet Allocation*, mais especificamente a implementação denominada JGibbsLDA[2], uma implementação em Java do GibbsLDA++ (PHAN; NGUYE, 2007), para obtenção dos tópicos futuramente empregados no enriquecimento do texto e em uma representação numérica.

4.3 Baseline TF-IDF

O baseline desse projeto replicou parte do trabalho proposto por Miyata, Kano e Digiampietri (2013), no sentido de usar a mesma abordagem que adota uma modificação do TF-IDF para fazer a classificação. O algoritmo TF-IDF foi implementado utilizando a linguagem Python. O cálculo de qual classe (área de atuação) um dado título do pesquisador pertence adotou a mesma abordagem de Miyata, Kano e Digiampietri (2013), dividindo-se a frequência de cada palavra no conjunto de dados da grande área (ou área ou subárea) avaliada em específico, pela frequência da mesma palavra no conjunto composto por todas as grande áreas (ou áreas ou subáreas) (MIYATA; KANO; DIGIAMPIETRI, 2013).

Seja w uma palavra presente dentro do documento atual d_j, c_i a grande área (ou área ou subárea) avaliada no momento, $c_i(w)$ a frequência que a palavra w tem dentro da classe c_i e $C(w)$ sendo a frequência dessa mesma palavra no conjunto de todas as classes, $P(c_i|d_j)$ é o valor de pertinência que o documento d_j tem para a classe (grande área, área ou subárea) c_i, e este é encontrado pela equação 6. A classe cujo $P(c_i|d_j)$ tiver o maior valor é dada como aquela que melhor representa a área de atuação de cada título do pesquisador.

$$P(c_i|d_j) = \sum_{w \in d_j} \frac{c_i(w)}{C(w)} \tag{6}$$

A classificação do pesquisador é feita da seguinte forma: para cada título da produção científica do pesquisador avaliado (títulos de artigos, orientações e projetos de pesquisa) uma classe é inferida usando a equação 6 e a classe que ganhar o voto majoritário dentre

[2] JGibbsLDA: ⟨http://jgibblda.sourceforge.net⟩, acessado em 30/01/2018

os seus títulos é tida como a classe do pesquisador em si. No caso de empate, uma das duas classes majoritárias é escolhida randomicamente.

4.4 TF-IDF como representação numérica

O TF-IDF é uma forma de representação que computa uma frequência ponderada das palavras nos textos para determinar a importância das mesmas. Usualmente se adota uma matriz termo-documento para representação do texto. Contudo, o tamanho do vocabulário influencia no tamanho da matriz e, assim, na complexidade do problema a ser tratado. O problema de inferência das áreas de atuação utilizando do Lattes possuí um vocabulário extenso, sobretudo por causa dos idiomas diversos que os títulos podem ser escritos, o que torna ineficiente empregar a matriz termo-documento diretamente.

Assim, o TF-IDF como classificador, adotado por Miyata, Kano e Digiampietri (2013), é modificado para ser utilizado como uma representação numérica que não depende da matriz termo-documento. Assim, ao invés de atribuir a classe com maior número de $P(c_i|d_j)$ como a correta, adota-se os valores calculados para todas as classes como entrada para outros algoritmos, transformando d_j em um vetor numérico $V(d_j) = [P(c_1|d_j), P(c_2|d_j), ..., P(c_i|d_j)]$.

Por exemplo, se o problema original tiver oito classes, transforma-se cada texto em uma matriz de oito números, reduzindo drasticamente o número de atributos do problema. Contudo, esse tipo de abordagem necessita de um conjunto de treinamento para conseguir aprender as funções $P(c_i|d_j)$ das classes em função do texto. Para isso, 90% do conjunto de dados original foi usado para fazer a representação por TF-IDF (treinamento) e o restante como entrada para os algoritmos de aprendizado de máquina (teste).

4.5 Enriquecimento do texto

O enriquecimento dos títulos foi feito usando uma abordagem semelhante a proposta por Vo e Ock (2015), que adota o algoritmo *Latent Dirichlet Allocation*. O número de tópicos no presente trabalho foi fixado em 150, assim como α e β do LDA fixados em $0,5$ e $0,01$, respectivamente, mantendo 200 palavras por tópico. Esse enriquecimento segue as seguintes fases:

1. Construção de um *corpus* de texto de uma base de dados externa (neste projeto, este corpus foi baseado na Wikipédia).
2. O *corpus* construído é utilizado como entrada para o LDA, o qual mapeia a distribuição de probabilidades das palavras em cada tópico.
3. Cada palavra presente nos títulos é mapeada para um conjunto de tópicos usando a distribuição obtida na fase anterior, adotando a equação 5, presente no capítulo 3.
4. Um ranking sobre os tópicos mapeados é feito para cada palavra, atribuindo como tópico principal aquele mais relevante dado o contexto do título.
5. Para cada palavra do título, as três palavras mais relevantes do seu tópico principal são adicionadas ao título, de forma a enriquecer o texto.

A fase de avaliação dos classificadores utilizou dos títulos enriquecidos para fazer o treinamento dos modelos adotados. Tais modelos gerados foram usados em alguns testes para verificar o desempenho da abordagem proposta.

4.6 Representação numérica baseada em tópicos

A representação numérica baseada em tópico considera os valores obtidos de cada palavra sobre os tópicos atribuídos a mesma como uma função de pertinência dela para os tópicos. Ela é uma abordagem que mistura as ideias propostas por Gabrilovich e Markovitch (2007) com o enriquecimento proposto por Vo e Ock (2015), na qual ao invés de usar os tópicos obtidos pelo LDA para enriquecer o texto, adota-se os valores obtidos neles como uma nova representação numérica para o texto.

Desta forma, para cada palavra em um dado texto, a probabilidade dos tópicos atribuídos as palavras é adicionada a probabilidade de tais tópicos a outras palavras no mesmo documento, de forma a seguir a equação 5. Além disso, nesta dissertação, o atributo β desta equação foi fixado em $0,5$, fazendo com que os outros tópicos adaptados tenham impacto no que está sendo calculado, porém evitando que esse impacto seja muito grande.

Considera-se os valores obtidos de cada palavra sobre os tópicos atribuídos a ela como uma função de pertinência dela para os mesmos, dentro do contexto de cada documento. Para cada tópico avaliado, seu valor correspondente em cada documento d_j é calculado usando a equação 7. A variável n representa o número de palavras presentes dentro de d_j. Assim, mapeia-se d_j em um vetor numérico $V(d_j) = [P(t_1|d_j), P(t_2|d_j), ..., P(t_i|d_j)]$,

que pode ser visto como uma nova representação para d_j. Esta representação tem como vantagem ser mais compacta e funcionar com diversos algoritmos de aprendizado de máquina.

$$P(t_i|d_j) = \frac{\sum_{w \in d_j} \overline{P_w(t_i)}}{n} \tag{7}$$

Com o intuito de avaliar o impacto do número de tópicos nessa representação, variou-se o seu valor em 150 e 300 tópicos, sempre com com α e β para obtenção dos tópicos no *Gibbs Sampling* fixados em $0,5$ e $0,01$, respectivamente e com 200 palavras em cada tópico para fazer o mapeamento.

4.7 Análise de Rede Social

Miyata, Kano e Digiampietri (2013) utilizam da estrutura em grafos (por meio das relações de coautoria entre os pesquisadores) da Plataforma Lattes para fazer uma Análise de Rede Social sobre os pesquisadores e assim tentar melhorar os resultados obtidos pelos algoritmos, mais especificamente, melhorar o desempenho do TF-IDF como classificador.

Duas características foram extraídas: a porcentagem dos vizinhos pertencentes a cada grande área (ou área ou subárea) utilizando o primeiro nível de vizinhos (V1 - apenas vizinhos diretos, ou seja, coautores) e a vizinhança de nível dois (V2 - vizinhos e vizinhos dos vizinhos) (MIYATA; KANO; DIGIAMPIETRI, 2013). No presente trabalho a mesma abordagem foi adotada, sendo que para se extrair essas características, três grafos de coautoria foram criados: contendo todas as publicações dos pesquisadores, publicações entre 2011 e 2015 (cinco últimos anos) e publicações entre 2013 e 2015 (três últimos anos). Esses três intervalos foram escolhidos para se adequar aos três intervalos de tempo avaliados.

Ambas características (representadas como porcentagens) são vistas como um novo vetor de características dos pesquisadores, podendo ser utilizadas em conjunto com a maioria das abordagens propostas neste trabalho. Avaliou-se o impacto que tais características tiveram nos resultados, bem como se elas melhoraram os mesmos ou não.

Para as abordagens baseadas no TF-IDF (tanto como classificador quanto a numérica), existe uma aplicação alternativa das métricas, na qual os dois vetores de características oriundos da Análise de Rede Social são utilizados da seguinte forma: soma-

se os valores resultantes do TF-IDF com um dos dois vetores (TF-IDF + V1 ou TF-IDF + V2) e os valores obtidos são utilizados para determinar qual classe tem o maior número de "pontos" (no caso de usar como classificador). Além da soma em si, duas informações adicionais foram passadas para os algoritmos: qual classe retornou a maior pontuação e o valor dessa pontuação, de forma a auxiliar os algoritmos na hora de escolher qual classe é a mais adequada.

Nota-se que para as abordagens numéricas, ao invés de somar as características da análise de rede social, pode-se utilizar ambas como características diretamente, ou seja, se o problema tem oito classes, com a análise de rede social combinada com os as características utilizando TF-IDF, por exemplo, representa-se o texto como dezesseis características (oito do TF-IDF e oito da rede social). Esta forma de representação também foi avaliada, tanto para a representação numérica usando TF-IDF, quanto para a baseada em tópicos.

4.8 Classificação em dois níveis

Os resultados dos testes feitos, sobretudo aqueles relacionados ao enriquecimento de texto (que serão apresentados e discutidos no capítulo 6), deram indícios de que existem dois grupos de classes dentro das Grandes Áreas: as classes que são relacionadas às disciplinas de ciências humanas e as que não são relacionadas.

A abordagem chamada de classificação hierárquica, que usualmente utiliza de uma estrutura pré-definida de classes ao invés de criar uma nova estrutura baseada na similaridade que classes têm umas com as outras, pode ser vista como um problema particular de classificação estruturada, na qual a saída do problema de classificação é definida sobre uma taxonomia de classes (JR.; FREITAS, 2011).

Duas formas de classificação hierárquica foram adotadas neste trabalho: *flat classification approach* e *local classifier approach*. *Flat classification approach* é a abordagem de classificação que foi adotada em todas as outras abordagens citadas anteriormente, que ignora completamente a hierarquia das classes predizendo somente classes folhas (JR.; FREITAS, 2011). Dessa forma, cada um dos três problemas de classificação tratados até então (Grandes Áreas, Áreas e Subáreas) pode ser considerado como uma classificação hierárquica usando *Flat classification approach*.

Por outro lado, *local classifier approach* leva em consideração a estrutura das classes, de forma *top-down* (classifica da raiz até as folhas) (JR.; FREITAS, 2011). Existem três formas de se fazer esse tipo de abordagem: classificação local por nó (*local classifier per node* - LCN, em inglês), classificação local por nó pai (*local classifier per parent node* - LCPN, em inglês) e classificação local por nível (*local classifier per level* - LCL, em inglês). Dentre essas, a abordagem LCL foi adotada neste trabalho.

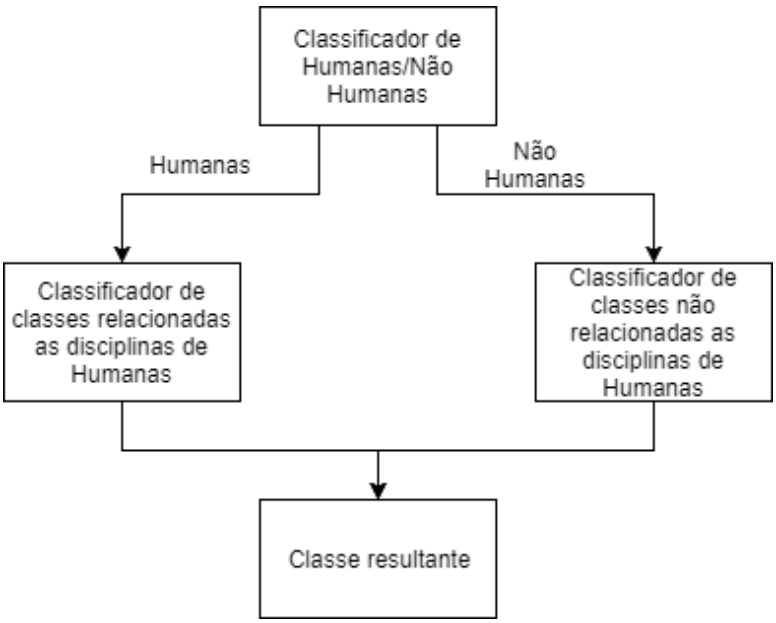
A classificação local por nível cria um classificador para cada nível da hierarquia (JR.; FREITAS, 2011). Por exemplo, no contexto do problema de pesquisa deste trabalho, seria fazer classificação hierárquica do problema de inferência das áreas de atuação dos pesquisadores, usando um classificador para cada nível (Grandes Áreas, Áreas e Subáreas) e utilizar a informação dos mesmos para dizer qual a estrutura da árvore de classes. Ou seja, considera-se cada nível como problemas separados (assim como este trabalho), mas o resultado de todos eles é analisado em conjunto, o que pode levar a inconsistências.

Por exemplo, um pesquisador pode ter sua Grande Área classificada como "Ciências Humanas", sua Área como "Agronomia" e Subárea classificada como "Computação Móvel". Desta forma, o resultado visto como um todo é inconsistente do ponto de vista da classificação hierárquica. Para evitar esse tipo de problema, neste trabalho em específico, considera-se cada nível como um problema separado e assim cada nível é avaliado separadamente.

A classificação hierárquica proposta nesse trabalho, chamada aqui de classificação em dois níveis por usar dois níveis de treinamento, difere do LCL comum ao utilizar de um novo problema criado artificialmente para ajudar na classificação. Esse novo problema avalia se o pesquisador pertence às disciplinas de humanas, ou se ele está em uma classe que não é relacionada às disciplinas de humanas. Três classificadores foram construídos: um para predizer se um pesquisador está relacionado a Humanas ou não, um classificador para as classes de Humanas e um classificador para as classes "Não-Humanas". A figura 4 mostra a estrutura do problema nesse contexto.

Este tipo de classificação só foi aplicada nas Grandes Áreas, uma vez que o número de classes é reduzido e este foi o problema que teve os resultados mais proeminentes no enriquecimento de texto (o que motivou adotar a classificação em dois níveis). Assim, as classes consideradas de Humanas foram "Ciências Humanas", "Linguística, Letras e Artes" e "Ciências Sociais e Aplicadas", enquanto que as classes ditas como Não-Humanas foram

Figura 4 – Estrutura da classificação em dois níveis



Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

"Ciências Agrárias", "Ciências Biológicas", "Ciências da Saúde", "Ciências Exatas e da Terra" e "Engenharias".

A construção do conjunto de treinamento para a classificação em dois níveis seguiu um 10-*fold cross validation* estratificado, considerando como se somente um classificador fosse treinado. Cada conjunto de treinamento é usado em sua totalidade para fazer o treinamento do primeiro nível da classificação, aquele que avalia se um pesquisador é das áreas de Humanas ou não.

Para fazer o treinamento do segundo nível de classificação, o conjunto de treinamento (o mesmo usado no primeiro nível) é dividido em duas parcelas da seguinte forma: instâncias que têm como classe aquelas relacionadas as classes de Humanas irão para o conjunto de treinamento do classificador de classes de Humanas e as instâncias restantes irão para o treinamento do classificador de classes Não-Humanas. Assim, cada um desses classificadores irá trabalhar a partir de um problema específico, para diferenciar em qual classe dentre as ditas de Humanas ou Não-Humanas o pesquisador atua.

O teste de uma nova instância segue uma abordagem *top-down* na hierarquia da figura 4, na qual se avalia, primeiramente, se o pesquisador é de Humanas ou Não-Humanas. Caso o pesquisador seja de humanas, o classificador específico para as classes relacionadas a humanas será então usado para tentar predizer qual classe dentre as de humanas esse pesquisador é classificado. A mesma coisa vale para pesquisadores ditos

como Não-Humanas, substituindo o classificador das classes dentre as de humanas para as classes dentre as de não-humanas. A avaliação de acurácia dessa abordagem é feita como usualmente: verificando se a classe resultante (depois dos dois níveis hierárquicos) é igual a esperada.

4.9 Uso da informação de hierarquia

Como no problema de inferência das áreas de atuação de pesquisadores existe uma taxonomia de classes, pode-se utilizar dessa estrutura para melhorar o resultado dos algoritmos, sem necessariamente depender de uma classificação hierárquica. Exemplificando, pode-se querer inferir qual Área um dado pesquisador possui, sabendo de antemão qual a Grande Área do mesmo. Assim, é possível usar o conhecimento prévio de qual classe o pesquisador possui no nível acima da taxonomia para tentar melhorar a inferência do nível atual. Tal conhecimento prévio pode ser útil por diminuir o tamanho do problema original, pois sabendo a classe do nível superior os algoritmos poderão dar mais enfase as classes mais próximas a ela, e assim acertar mais qual classe um pesquisador pertence.

Para avaliar qual o impacto que esse tipo de informação tem sobre os resultados da classificação, um novo campo denominado “classe superior” foi adicionado aos conjuntos de dados, possuindo o nome da classe do nível imediatamente superior ao avaliado desse pesquisador. Consequentemente, somente as Áreas e as Subáreas serão testadas com esse campo, uma vez que não existe um nível superior às Grandes Áreas na taxonomia de classes.

5 Testes

Neste capítulo, os testes feitos sobre as abordagens descritas no capítulo anterior são detalhados, bem como o que se pretende avaliar com cada um deles. Para a realização dos testes, um computador com processador i7 de 3.4 GHz, 12 Gb de memória RAM e sistema operacional Windows 10 foi utilizado. A princípio, um dos objetivos desse projeto é determinar se fazer o enriquecimento do texto dos títulos das produções científicas, usando a abordagem descrita por Vo e Ock (2015), melhora o desempenho da inferência das áreas de atuação dos pesquisadores. Assim, os testes iniciais têm o intuito de fazer esse tipo de avaliação.

As abordagens que seguem o *baseline* proposto por Miyata, Kano e Digiampietri (2013) e o TF-IDF como representação numérica serão as adotadas para avaliar o desempenho do enriquecimento, comparando seus resultados com os obtidos no conjunto de dados sem enriquecimento e com o conjunto traduzido.

Os três níveis da taxonomia (Grandes Áreas, Áreas e Subáreas) tiveram seus resultados computados nesse teste (em relação à acurácia), contudo, matrizes de confusão das Grandes Áreas serão apresentadas também para indicar o quanto o enriquecimento influencia no desempenho classe a classe.

O próximo passo foi comparar o desempenho das representações numéricas entre si (TF-IDF contra baseada em tópicos), para determinar qual das duas possui os melhores resultados nos três problemas tratados (inferência das grandes áreas, áreas e subáreas).

Além disso, assim como feito por Miyata, Kano e Digiampietri (2013), a estrutura inerente de rede social foi utilizada para a extração de dois novos vetores de características, denominados V1 e V2. Tais foram testados junto do TF-IDF como classificador e para as duas representações numéricas avaliadas pelo projeto.

Outro teste realizado se refere a tradução dos títulos, mais especificamente, o quanto traduzir o texto influencia ou não no desempenho dos algoritmos. Neste caso, comparou-se os resultados obtidos nos textos com e sem tradução em todas as abordagens avaliadas. A porcentagem de texto em determinadas línguas também pode influenciar no desempenho, assim, computou-se as porcentagem de títulos em português, inglês, espanhol e outras línguas e esses quatro valores foram utilizados como novas características e adicionados às presentes para avaliar se há melhora no desempenho dos algoritmos.

Em suma, para cada um dos testes feitos um conjunto de dados foi construído. Uma descrição de cada um deles está presente no Quadro 3. Cada um dos conjuntos descritos possuem variantes que consideram os três problemas a serem resolvidos (Grandes Áreas, Áreas e Subáreas), o fato de terem sido traduzidos ou enriquecidos e os três intervalos de tempo (toda produção científica, últimos cinco anos, últimos três anos).

Quadro 3 – Descrição dos conjuntos construídos para os Testes

Conjunto	**Descrição**
LattesDB	Conjunto de dados original obtido do Lattes (texto)
TfIdfDB	Representação numérica do Lattes usando o TF-IDF
LanguageDB	Representação baseada na proporção dos idiomas dos títulos
VizinhançaXDB	Representação que utiliza das vizinhanças de Nível X para representar os títulos*
TfIdf+VXDB	Representação numérica do Lattes usando o TF-IDF somada aos valores da vizinhança de nível X*
Language-TfIdfDB	Representação numérica do Lattes usando o TF-IDF e a proporção dos idiomas dos títulos
Language-TfIdfVXDB	Representação numérica do Lattes, proporção dos idiomas dos títulos e a vizinhança de nível X*
Wiki-X	Representação numérica do Lattes, usando um número X de tópicos obtidos do Wikipédia*
Wiki-VX-Y	Representação numérica do Lattes, usando um número Y de tópicos obtidos do Wikipédia e a vizinhança de nível X*
HierarquicoDB	Representação numérica do Lattes por meio do TF-IDF com três subconjuntos para a classificação em dois níveis
HierarquicoWiki-X	Representação numérica do Lattes usando X tópicos com três subconjuntos para a classificação em dois níveis*
HierarquiaClasseDB	Representação numérica do Lattes por meio do TF-IDF adicionados a informação da hierarquia de classe
HierarquiaClasseWiki-X	Representação numérica do Lattes usando X tópicos adicionados a informação da hierarquia de classe*

* Neste trabalho foram utilizadas vizinhança de nível 1 e 2 (X=1 e X=2) e número de tópicos igual a 150 e 300 (Y=150 e Y=300).

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Os conjuntos LattesDB, Wiki-X, Wiki-VXY, HierarquicoWiki-X e HierarquiaClasseWiki-X adotam o conjunto completo para a realização dos testes utilizando validação cruzada com dez subconjuntos *10-fold cross validation*, enquanto que os conjuntos restantes utilizam de 90% do conjunto para realizar o treinamento das técnicas usadas para a construção dos mesmos e os 10% restantes são adotados para a representação numérica e subsequentemente serem avaliadas utilizando a validação cruzada.

O conjunto LattesDB é aquele que tem como valores para cada pesquisador o texto de sua produção científica. Desta forma, esse conjunto tem como atributos as palavras presentes nos títulos, e estas são usadas como entrada para a abordagem *baseline* proposta por Miyata, Kano e Digiampietri (2013), que usa uma abordagem parecida com o TF-IDF para fazer a classificação dos pesquisadores.

TfidfDB é o conjunto que usa abordagem de representação numérica por TF-IDF para transformar o texto da produção científica dos pesquisadores em um vetor de números, sendo 8 valores para as Grandes Áreas, 75 para as Áreas e 484 para as Subáreas, o mesmo número de classes de cada problema. Tal conjunto é menor do que o LattesDB, pois 90% do conjunto original foi utilizado para treinamento da abordagem, sendo o tamanho do mesmo então apenas 10% do tamanho do conjunto original.

O TfidfDB possui algumas variantes denominadas TfIdf+VXDB, Language-TfIdfDB e Language-TfIdfVXDB, que adicionam novas características ao conjunto original de acordo com a abordagem testada. Por exemplo, o TfIdf+VXDB continua possuindo somente 8, 75 ou 484 atributos (a depender do nível tratado), pois os valores calculados para a métrica de rede são somados diretamente aos seus correspondentes do TF-IDF. Enquanto que o conjunto Language-TfIdfDB possuí 12, 79 ou 488 atributos de entrada (a depender do nível tratado), porque 4 novas características de proporção de idioma foram adicionadas ao mesmo. Por fim, o conjunto Language-TfIdfVXDB possuí 20, 154 ou 972 características, porque ao invés de somar diretamente os valores das métricas de rede social ao TF-IDF, adota-se as mesmas como atributos novos.

Wiki-X é o conjunto que usa a abordagem de representação numérica por tópicos para transformar o texto da produção científica dos pesquisadores em um vetor de X números, sendo X o número de tópicos extraídos da base de dados externa. Possuí a variante Wiki-VX-Y, que além de usar os Y tópicos (nesse caso) para representar o texto, também usa as métricas de rede social de vizinha um ou dois como atributo de entrada.

Os conjuntos HierarquicoDB e HierarquicoWiki-X são idênticos as contrapartes TfidfDB e Wiki-X, porém foram separados em subconjuntos para fazer o treinamento dos algoritmos usados na classificação em dois níveis. Igualmente, os conjuntos HierarquiaClasseDB e HierarquiaClasseWiki-X são idênticos as suas contrapartes TfidfDB e Wiki-X, somente possuindo como adicional o atributo que é a classe do nível acima do pesquisador.

6 Resultados e Discussões

Neste capítulo os resultados obtidos por meio dos testes descritos no capítulo 5 são apresentados, bem como as devidas considerações sobre os mesmos. Para efeito de organização eles são divididos em seções que apresentam o objetivo de cada teste feito, quais resultados eram esperados e uma comparação entre eles.

O *baseline* do projeto, resultados que seguem a mesma abordagem proposta por Miyata, Kano e Digiampietri (2013), é discutido na seção 6.1. A seção 6.2 apresenta os resultados oriundos do enriquecimento de texto, mostrando se enriquecer melhora ou não o desempenho e porque. A seção 6.3 discute os resultados das abordagens numéricas no geral, comparando-as e delineando quais possuem os melhores resultados para os três problemas tratados nesse projeto de pesquisa (Grandes Áreas, Áreas e Subáreas). O uso da proporção dos idiomas nos títulos é avaliado na seção 6.4. A seção 6.5 analisa as contribuições que a análise de rede social tem sobre os resultados das representações como um todo. A seção 6.6 apresenta os resultados da classificação em dois níveis, de forma a tentar melhorar as acurácias encontradas até então. A aplicação da informação sobre a hierarquia das classes é avaliada na seção 6.7, de forma a determinar a influência desta. Na seção 6.8 são apresentados os testes estatísticos comparando os melhores resultados obtidos para encontrar qual de todas as abordagens propostas tem o melhor resultado no geral.

6.1 Baseline

Os resultados do *baseline* seguem a mesma abordagem de Miyata, Kano e Digiampietri (2013), somente considerando a parte de mineração de texto e podem ser consultados nas tabelas 2, 3 e 4. Todos esses testes utilizaram do TF-IDF como classificador adotando o 10-*fold cross validation* estratificado para computar a acurácia média da abordagem.

O melhor resultado obtido para as Grandes Áreas por Miyata, Kano e Digiampietri (2013), usando somente o TF-IDF como mineração de texto, obteve $83,670\%$ de acurácia com 9 anos da produção científica dos pesquisadores. Como é observável na tabela 2, os resultados encontrados nesta pesquisa foram superiores aos encontrados no trabalho original, mesmo utilizando uma janela de tempo menor.

Tabela 2 – Resultados do Baseline - Grandes Áreas

Período analisado	Conjunto	Acurácia(%)	Desvio Padrão(%)
Últimos 3 Anos	Normal	88,07	1,3
	Traduzido	87,36	1,2
Últimos 5 Anos	Normal	88,83	0,8
	Traduzido	88,80	0,9
Produção Completa	Normal	90,84	0,9
	Traduzido	90,77	0,8

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Provavelmente tal fato ocorre porque a base de dados desse trabalho não é a mesma de Miyata, Kano e Digiampietri (2013), uma vez que ela possui dados mais atualizados. Além disso, os títulos da produção científica desse trabalho consideram os títulos dos artigos, projetos de pesquisa e orientações, enquanto o trabalho de Miyata, Kano e Digiampietri (2013) considera somente o títulos dos artigos, o que potencialmente faz com que o conjunto resultante não seja tão representativo quanto o adotado neste projeto.

Tabela 3 – Resultados do Baseline - Áreas

Período analisado	Conjunto	Acurácia(%)	Desvio Padrão(%)
Últimos 3 Anos	Normal	70,91	1,7
	Traduzido	68,36	1,2
Últimos 5 Anos	Normal	75,19	1,2
	Traduzido	73,20	1,9
Produção Completa	Normal	82,11	1,3
	Traduzido	80,57	1,2

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Da mesma forma, os resultados para as Áreas estão presentes na tabela 3. O melhor resultado usando mineração de dados, originalmente, era uma acurácia de $68,220\%$ com 9 anos de produção científica sendo considerados. Como ocorreu com as Grandes Áreas, os resultados obtidos para o problema das Áreas foram superiores aos obtidos no trabalho original, possuindo os mesmo motivos apresentados anteriormente.

Por fim, os resultados referentes as Subáreas estão presentes na tabela 4. O melhor resultado obtido por Miyata, Kano e Digiampietri (2013) foi de $26,530\%$ com 9 anos de produção científica. Igualmente aos resultados anteriores, mesmo com uma janela de tempo menor (últimos 5 anos) a presente pesquisa obteve resultados comparáveis aos do artigo original, o que fornece indícios de que utilizar os títulos dos projetos de pesquisa e

Tabela 4 – Resultados do Baseline - Subáreas

Período analisado	Conjunto	Acurácia(%)	Desvio Padrão(%)
Últimos 3 Anos	Normal	21,25	1,7
	Traduzido	20,64	2,4
Últimos 5 Anos	Normal	26,27	2,2
	Traduzido	25,02	2,5
Produção Completa	Normal	32,58	3,2
	Traduzido	31,47	2,4

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

das orientações, juntos dos títulos dos artigos publicados têm resultados melhores do que somente usar os títulos dos artigos.

As acurácias obtidas, no geral, mostram que existe uma tendência de que quanto maior for a janela de tempo para a produção científica considerada para o TF-IDF, melhor será o desempenho da abordagem. Isso fornece indícios de que os pesquisadores não mudam muito suas áreas de atuação com o passar do tempo, contudo uma análise mais detalhada é necessária, uma vez que também existe a possibilidade dos resultados melhorarem somente porque o algoritmo possui mais dados para trabalhar.

Outro ponto que vale a pena ser analisado é a contribuição de fazer a tradução do texto ou não. Pelo menos no que diz respeito ao *baseline*, traduzir o texto faz com que a abordagem perca, na média, cerca de 1% de acurácia. Isto indica que, no Brasil, algumas áreas publicam mais em alguns idiomas do que outras e isto pode ser utilizado pelos sistemas de inferência de área de atuação.

Os melhores valores presentes no *baseline*, para cada uma das classes, são comparados a cada um dos resultados encontrados nas outras abordagens, de forma a apresentar a diferença entre o valor encontrado e aquele do *baseline*. Assim, nas tabelas desse capítulo os valores entre parenteses irão representar a diferença entre o valor apresentado na tabela e aquele correspondente para o mesmo problema e mesma janela de tempo. Por exemplo, caso o resultado seja referente as Áreas usando toda produção científica, o valor entre parenteses irá mostrar a diferença entre o valor encontrado e $82,11\%$, o melhor valor encontrado para as Áreas usando toda produção científica no *baseline*.

Os valores comparados então são $88,07\%$, $88,83\%$ e $90,84\%$ para 3 anos de produção, 5 anos de produção e toda produção científica, respectivamente, nas Grandes Áreas. Da mesma forma, $70,91\%$, $75,19\%$ e $82,11\%$ representam 3 anos de produção, 5 anos de produção e toda produção científica, para as Áreas. Por fim, $21,25\%$, $26,27\%$ e $32,58\%$

são os valores comparados nas Subáreas, para 3 anos de produção, 5 anos de produção e toda produção científica, respectivamente.

6.2 Enriquecimento de Texto

Os resultados oriundos do enriquecimento do texto foram obtidos a partir dos conjuntos LattesDB (o mesmo do *baseline*) e TfidfDB, devendo ser comparados aos encontrados na seção 6.1 para verificar se enriquecer o texto melhora ou não o desempenho da tarefa de inferência. Todos os testes utilizaram 10-*fold cross validation* estratificado para computar a acurácia média da abordagem.

Tabela 5 – Resultados do Enriquecimento - LattesDB

Período analisado	**Conjunto**	**Acurácia(%)**	**Desvio Padrão(%)**
Últimos 3 Anos	Grandes Áreas	81,28(-6,79)	1,2
	Áreas	64,87(-6,04)	1,4
Últimos 5 Anos	Grandes Áreas	82,17(-6,66)	1,4
	Áreas	68,95(-6,24)	2,0
Produção Completa	Grandes Áreas	84,02(-6,82)	1,2
	Áreas	76,31(-5,8)	1,7

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

A tabela 5 mostra resultados obtidos com base no conjunto LattesDB e usando a abordagem de Miyata, Kano e Digiampietri (2013) para fazer a classificação dos pesquisadores com os títulos das produções científicas enriquecidos. Nota-se que não foram computados resultados para as Subáreas, uma vez constatado que o enriquecimento potencialmente piora o desempenho, como é possível se perceber nos valores entre parênteses (diferença com os resultados do baseline), que por serem negativos mostram que o enriquecimento piorou a acurácia no geral.

Contudo, com o intuito de analisar qual a influência do enriquecimento sobre algoritmos diversos também se computou os resultados gerais com os conjuntos TfidfDB para dois dos problemas avaliados (Grandes Áreas e Áreas). Tais resultados estão presentes nas tabelas 6 e 7, nas quais se confirmou que o enriquecimento dos títulos da maneira que foi realizado e considerando os problemas tratados piorou o desempenho dos algoritmos para o problema da inferência das áreas de atuação de pesquisadores.

Tabela 6 – Resultados do conjunto TfidfDB (Grandes Áreas) - Acurácia (%)

Período analisado	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
3 Anos	Normal	88,35(+0,28)	85,04(-3,03)	90,27(+2,2)	90,91(+2,84)
	Traduzido	85,77(-2,3)	85,88(-2,19)	88,34(+0,27)	89,19(+1,12)
	Enriquecido	76,25(-11,8)	68,02(-20,1)	82,35(-5,72)	79,03(-9,04)
5 Anos	Normal	86,96(-1,87)	86,75(-2,08)	89,53(+0,7)	90,27(+1,44)
	Traduzido	90,38(+1,55)	86,64(-2,19)	91,13(+2,3)	91,34(+2,51)
	Enriquecido	79,06(-9,77)	69,23(-19,6)	83,01(-5,82)	84,72(-4,11)
Completo	Normal	90,27(-0,57)	89,63(-1,21)	91,02(+0,18)	92,20(+1,36)
	Traduzido	88,78(-2,06)	88,78(-2,06)	91,34(+0,5)	91,23(+0,39)
	Enriquecido	82,05(-8,79)	75,85(-14,9)	87,07(-3,77)	88,78(-2,06)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Tabela 7 – Resultados do conjunto TfidfDB (Áreas) - Acurácia (%)

Período analisado	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
3 Anos	Normal	53,30(-17,6)	56,53(-14,4)	57,26(-13,6)	74,30(+3,39)
	Traduzido	48,45(-22,4)	56,55(-14,3)	58,17(-12,7)	71,42(+0,51)
	Enriquecido	48,67(-22,2)	42,64(-28,2)	50,00(-20,9)	65,58(-5,33)
5 Anos	Normal	56,38(-18,8)	60,35(-14,8)	61,67(-13,5)	79,29(+4,1)
	Traduzido	55,06(-20,1)	59,61(-15,5)	59,03(-16,1)	77,53(+2,34)
	Enriquecido	46,10(-29,1)	44,20(-30,9)	53,59(-21,6)	69,75(-5,44)
Completo	Normal	63,28(-18,8)	64,46(-17,6)	66,96(-15,1)	81,79(-0,32)
	Traduzido	60,64(-21,4)	67,10(-15,0)	64,61(-17,5)	80,32(-1,79)
	Enriquecido	54,77(-27,3)	52,27(-29,8)	60,20(-21,9)	74,89(-7,22)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Assim sendo, torna-se necessário tentar compreender porque o enriquecimento piorou o desempenho da maneira como foi verificado. O enriquecimento proposto por Vo e Ock (2015) foi avaliado originalmente a partir de um problema com classes disjuntas, ou seja, classes que não possuem sobreposição umas sobre as outras.

Todavia, como os três problemas tratados nesse projeto envolvem classes com sobreposição ou com fronteiras tênues, uma das considerações feitas é que talvez o enriquecimento esteja adicionando ambiguidade ao texto dos títulos, de forma a dificultar a capacidade dos algoritmos em distinguirem classes muito próximas. Para verificar essa hipótese, torna-se necessário avaliar as matrizes de confusão dos resultados para ver o que está acontecendo, sendo estas presentes nas tabelas 8, 9, 10 e 11.

O quadro 4 apresenta a definição do que cada letra presente nas matrizes de confusão significa, ou seja, qual classe cada linha representa. Apenas as Grandes Áreas estão sendo

Quadro 4 – Descrição das classes usadas na matriz de confusão

Letra	**Classe**
A	Ciências Agrárias
B	Ciências Biológicas
C	Ciências da Saúde
D	Ciências Exatas e da Terra
E	Ciências Humanas
F	Ciências Sociais e Aplicadas
G	Engenharias
H	Linguística, Letras e Artes

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

consideradas nessas matrizes e toda produção científica dos pesquisadores foi utilizada (no treinamento ou nos testes).

Tabela 8 – Matriz de Confusão do Texto Original - Grandes Áreas (LattesDB)

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	99	1	1	1	0	1	0	0
B	0	159	1	0	0	0	0	0
C	0	6	95	1	1	2	0	1
D	0	5	0	203	0	0	1	1
E	0	0	2	0	114	8	0	14
F	0	0	0	1	4	51	1	5
G	0	0	1	9	0	6	98	0
H	0	0	0	0	0	0	0	43

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Um ponto interessante das matrizes nas tabelas 8 e 9 está no fato de que a classe E (Ciências Humanas) se confunde muito com as classes F (Ciências Sociais e Aplicadas) e H (Linguística, Letras e Artes), sendo isso explicável uma vez que as três classes possuem uma certa sobreposição sobre o que elas representam. Contudo, a matriz referente ao enriquecimento (tabela 9) mostra que ocorreu um aumento considerável da confusão entre a classe E (Ciências Humanas) e a classe H (Linguística, Letras e Artes), na qual cerca 30% de todos os pesquisadores da classe E são confundidos como pertencendo à classe H pelo classificador por TF-IDF. Isso fornece evidências de que o enriquecimento utilizado adiciona ambiguidade no texto, o que faz o resultado no geral piorar.

Da mesma forma, as tabelas 10 e 11 apresentam as matrizes de confusão obtidas usando o conjunto TfidfDB (ou seja, a representação numérica do TF-IDF) usando *Support Vector Machine* como abordagem. Neste caso, os resultados com o texto enriquecido

Tabela 9 – Matriz de Confusão do Texto Enriquecido - Grandes Áreas (LattesDB)

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	93	5	5	0	0	0	0	0
B	5	150	5	0	0	0	0	0
C	1	6	95	2	0	0	2	0
D	3	8	1	186	0	3	7	2
E	0	0	2	0	86	9	0	41
F	0	0	0	0	7	50	0	5
G	2	0	0	5	0	4	103	0
H	0	0	0	0	0	0	0	43

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

(tabela 11) mostram que a SVM confundiu mais a classe F (Ciências Sociais e Aplicadas) com a classe E (Ciências Humanas), além de também ter ocorrido confusão da classe G (Engenharias) com a classe D (Ciências Exatas e da Terra).

No geral, as classes confundidas tem sobreposição umas com as outras, o que novamente é uma evidência de que o enriquecimento piora o resultado da inferência por adicionar ambiguidade ao texto.

Tabela 10 – Matriz de Confusão do Texto Original - Grandes Áreas (TfidfDB)

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	95	3	4	0	0	0	1	0
B	6	147	5	2	0	0	0	0
C	0	4	98	3	1	0	0	0
D	2	5	0	198	1	0	4	0
E	0	0	1	1	131	4	0	1
F	0	0	1	0	15	46	0	0
G	0	0	0	1	0	2	111	0
H	0	0	0	0	6	0	0	37

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

6.3 Abordagens Numéricas

Nesta seção os resultados das abordagens numéricas, seja a baseada no TF-IDF (TfidfDB), quanto aquela baseada na representação por tópicos (Wiki-X) são discutidos com maiores detalhes. Parte deles, mais especificamente os resultados do conjunto TfidfDB para as Grandes Áreas e Áreas, já foram apresentados nas tabelas 6 e 7. A tabela 12 mostra os resultados dessa representação para o problema das Subáreas.

Tabela 11 – Matriz de Confusão do Texto Enriquecido - Grandes Áreas (TfidfDB)

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	94	5	3	0	0	0	1	0
B	7	151	2	0	0	0	0	0
C	0	6	98	0	1	0	1	0
D	6	4	0	191	0	0	9	0
E	0	0	3	0	131	3	0	1
F	0	0	0	0	24	37	1	0
G	2	0	0	14	0	3	95	0
H	0	0	0	0	9	0	0	34

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Tabela 12 – Resultados do conjunto TfidfDB (Subáreas) - Acurácia (%)

Período analisado	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
3 Anos	Normal	16,94(-4,31)	22,59(+1,34)	26,83(+5,58)	44,06(+22,8)
	Traduzido	16,43(-4,82)	24,07(+2,82)	26,06(+4,81)	40,51(+19,2)
5 Anos	Normal	14,97(-11,3)	25,98(-0,29)	28,53(+2,26)	44,91(+18,6)
	Traduzido	16,43(-9,84)	24,92(-1,35)	27,19(+0,92)	43,90(+17,6)
Completo	Normal	25,42(-7,16)	29,94(-2,64)	37,00(+4,42)	51,41(+18,8)
	Traduzido	22,31(-10,2)	28,81(-3,77)	33,61(+1,03)	52,26(+19,6)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

No geral, o algoritmo que obteve o melhor desempenho dentre os adotados no conjunto TfidfDB foi a *Support Vector Machine* (SVM), que retornou a melhor acurácia em todos os testes feitos. Além disso, tal abordagem alcançou acurácias maiores que o *baseline*.

Por exemplo, as melhores acurácias para os três problemas tratados (Grandes Áreas, Áreas e Subáreas) foram, respectivamente, $92,20\%$, $81,79\%$ e $51,41\%$, sendo que as obtidas no *baseline* para os mesmos problemas foram $90,84\%$, $82,11\%$ e $26,27\%$. Ou seja, os resultados para Grandes Áreas e Subáreas foram melhores (sendo que o obtido nas Subáreas foi praticamente duas vezes melhor), enquanto que a Área teve um decréscimo de 1% na acurácia. Isso fornece evidências de que a abordagem numérica por TF-IDF é melhor do que somente utilizar do TF-IDF como classificador.

Ademais, resultados similares foram encontrados usando a abordagem numérica por meio de tópicos, proposta nesta dissertação, denominada aqui pelos conjuntos Wiki-X. O X presente no nome representa o número de tópicos adotados para representar os testes,

que no caso deste trabalho são fixados em 150 e 300. Assim, construiu-se dois conjuntos, denominados Wiki-150 e Wiki-300, para os algoritmos serem adotados.

Até então, como os melhores resultados foram aqueles que utilizam de toda produção científica para a representação do TfidfDB, para os testes subsequentes (representação por tópicos) não se separou a produção em janelas de tempos, considerando então toda produção científica do pesquisador. As tabelas 13 e 14 apresentam os resultados dessa abordagem.

Tabela 13 – Resultados do conjunto Wiki-150 (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	74,58(-16,2)	70,75(-20,1)	84,37(-6,47)	87,57(-3,27)
	Traduzido	76,36(-14,4)	76,51(-14,3)	86,70(-4,14)	90,10(-0,74)
Área	Normal	52,85(-29,2)	54,75(-27,3)	67,53(-14,5)	70,50(-11,6)
	Traduzido	55,80(-26,3)	65,88(-16,2)	73,16(-8,95)	82,24(+0,13)
Subárea	Normal	22,90(-9,68)	33,81(+1,23)	41,25(+8,67)	39,63(+7,05)
	Traduzido	27,80(-4,78)	46,42(+13,8)	48,17(+15,5)	53,44(+20,8)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Dois pontos são destaque na representação numérica baseada em tópicos: novamente a SVM foi o algoritmo que obteve os melhores resultados dentre todos os testados. Além disso, os resultados melhoram substancialmente quando o conjunto de dados é baseado nos títulos traduzidos, o que vai contra a tendência vista anteriormente, cujos resultados pioravam.

Tal comportamento pode ser explicado uma vez que o conjunto de dados externo possui texto majoritariamente em português, fazendo os tópicos encontrados pelo LDA possuírem mais palavras em português. Assim, um conjunto de dados de texto misto (em vários idiomas) como os títulos originais presentes no Lattes é inadequado para fazer um mapeamento direto, sem nenhum tipo de tratamento no texto. Por este motivo a tradução melhora os resultados dessa representação.

Ao comparar os dois conjuntos Wiki-150 e Wiki-300, observa-se que as acurácias obtidas pelo conjunto Wiki-300 foram melhores que as do conjunto Wiki-150, o que dá evidências que quanto maior o número de tópicos adotados na representação, melhores são os resultados. Porém, identificar o número ideal de tópicos é uma atividade que ainda precisaria ser feita em um trabalho futuro.

Tabela 14 – Resultados do conjunto Wiki-300 (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	76,08(-14,7)	74,97(-15,8)	85,38(-5,46)	89,91(-0,93)
	Traduzido	77,35(-13,4)	80,96(-9,88)	87,25(-3,59)	92,09(+1,25)
Área	Normal	53,59(-28,5)	58,24(-23,8)	67,65(-14,4)	76,69(-5,42)
	Traduzido	58,23(-23,8)	69,71(-12,4)	73,17(-8,94)	85,86(+3,75)
Subárea	Normal	23,52(-9,06)	36,57(+3,99)	40,98(+8,4)	47,11(+14,5)
	Traduzido	28,96(-3,62)	47,53(+14,9)	48,05(+15,4)	59,48(+26,9)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Outro ponto se refere aos resultados da representação numérica baseada em TF-IDF comparada com a baseada em tópicos, no sentido de tentar verificar quais das duas tem os melhores resultados. Considerando somente o algoritmo *Support Vector Machine* (SVM) e os melhores resultados do TfidfDB e do Wiki-300 com toda produção científica (sem intervalo de tempo), nota-se que os resultados da representação baseada em tópicos teve acurácias superiores a representação baseada no TF-IDF, o que fornece indícios de que tal representação é superior.

Contudo, testes estatísticos são necessários para se fazer qualquer tipo de afirmação mais incisiva sobre qual das duas representações realmente possui o melhor desempenho na tarefa de inferência das áreas de atuação dos pesquisadores, uma vez que a acurácia pode ser considerada uma métrica fraca de comparação. Tais testes estatísticos são feitos na seção 6.8.

6.4 Uso da proporção dos idiomas

Um estudo possível envolve considerar a proporção dos idiomas dos títulos da produção científica de cada pesquisador, junto das abordagens estudadas anteriormente, de forma a avaliar se essa informação trás alguma melhoria para os algoritmos ou não. Para tal, o desempenho dos algoritmos nos conjuntos LanguageDB e Language-TfidfDB são avaliados.

Os resultados do conjunto LanguageDB estão presentes na Tabela 15, cuja acurácia mostra que a proporção dos idiomas não é bem representativa, no sentido de não conseguir representar bem cada pesquisador, o que explica o porque o desempenho dos algoritmos ser consideravelmente inferior aos obtidos até então.

Tabela 15 – Resultados do conjunto LanguageDB (toda produção científica) - Acurácia (%)

Nível	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	42,20(-48,6)	44,58(-46,2)	42,75(-48,0)	45,47(-45,3)
Área	16,57(-65,5)	22,23(-59,8)	17,21(-64,9)	22,23(-59,8)
Subárea	7,29(-25,2)	7,61(-24,9)	7,36(-25,2)	11,77(-20,8)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Tabela 16 – Resultados do conjunto Language-TfidfDB (Grandes Áreas) - Acurácia (%)

Período analisado	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
3 Anos	Normal	88,24(+0,17)	84,72(-3,35)	90,17(+2,1)	89,95(+1,88)
	Traduzido	84,59(-3,48)	84,49(-3,58)	88,77(+0,7)	88,87(+0,8)
5 Anos	Normal	85,36(-3,47)	83,54(-5,29)	90,91(+2,08)	89,95(+1,12)
	Traduzido	90,27(+1,44)	84,93(-3,9)	91,66(+2,83)	90,49(+1,66)
Completo	Normal	88,99(-1,85)	86,75(-4,09)	90,59(-0,25)	91,45(+0,61)
	Traduzido	88,78(-2,06)	88,78(-2,06)	91,34(+0,5)	91,23(+0,39)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Tabela 17 – Resultados do conjunto Language-TfidfDB (Áreas) - Acurácia (%)

Período analisado	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
3 Anos	Normal	54,03(-16,8)	56,82(-14,0)	58,00(-12,9)	74,30(+3,39)
	Traduzido	49,04(-21,8)	56,70(-14,2)	57,14(-13,7)	72,60(+1,69)
5 Anos	Normal	56,09(-19,1)	60,05(-15,1)	59,61(-15,5)	80,02(+4,83)
	Traduzido	52,86(-22,3)	61,08(-14,1)	59,91(-15,2)	79,00(+3,81)
Completo	Normal	63,14(-18,9)	63,87(-18,2)	66,96(-15,1)	81,05(-1,06)
	Traduzido	61,23(-20,8)	67,10(-15,0)	66,07(-16,0)	79,88(-2,23)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

As tabelas 16, 17 e 18 mostram os resultados a partir dos conjuntos Language-TfidfDB para cada um dos três problemas (Grandes Áreas, Áreas e Subáreas). Comparando-os com os equivalentes dos conjuntos TfidfDB (tabelas 6, 7 e 12), presentes nas seções 6.2 e 6.3, observa-se que utilizar a proporção dos idiomas é irrelevante para os algoritmos, uma vez que as acurácias ficaram muito próximas.

Tabela 18 – Resultados do conjunto Language-TfidfDB (Subáreas) - Acurácia (%)

Período analisado	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
3 Anos	Normal	14,12(-7,13)	23,16(+1,91)	29,09(+7,84)	44,91(+23,6)
	Traduzido	17,84(-3,41)	24,07(+2,82)	24,92(+3,67)	40,51(+19,2)
5 Anos	Normal	16,94(-9,33)	26,27(0)	27,11(+0,84)	45,76(+19,4)
	Traduzido	17,28(-8,99)	24,64(-1,63)	28,04(+1,77)	43,34(+17,0)
Completo	Normal	25,70(-6,88)	29,94(-2,64)	36,15(+3,57)	51,13(+18,5)
	Traduzido	28,81(-3,77)	28,81(-3,77)	33,61(+1,03)	52,54(+19,9)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

6.5 Análise de Rede Social

Os resultados sobre os testes envolvendo a técnica de Análise de Rede Social, adotada por Miyata, Kano e Digiampietri (2013), são aqueles obtidos a partir dos conjuntos VizinhançaXDB, Tfidf+VXDB, Wiki-VX-Y e Language-TfidfVXDB, sendo X em todas as nomenclaturas o tamanho da vizinhança considerada no cálculo das métricas, V1 os chamados vizinhos diretos (nível um) e V2 vizinhos diretos (nível um) e vizinhos dos vizinhos (nível dois).

Tabela 19 – Resultados dos conjuntos Vizinhanca1DB e Vizinhaca2DB - Texto original (Grande Áreas)

Período analisado	**V1**	**V2**
Últimos 3 Anos	87,49(-0,58)	86,43(-1,64)
Últimos 5 Anos	88,83(0)	87,12(-1,71)
Produção Completa	92,68(+1,84)	88,79(-2,05)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Tabela 20 – Resultados dos conjuntos Vizinhanca1DB e Vizinhaca2DB - Texto original (Áreas)

Período analisado	**V1**	**V2**
Últimos 3 Anos	75,41(+4,5)	74,23(+3,32)
Últimos 5 Anos	79,88(+4,69)	76,32(+1,13)
Produção Completa	85,91(+3,8)	79,99(-2,12)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

As tabelas 19, 20 e 21 mostram os resultados encontrados nos conjuntos Vizinhanca1DB e Vizinhanca2DB, seguindo a abordagem de Miyata, Kano e Digiampietri

Tabela 21 – Resultados dos conjuntos Vizinhanca1DB e Vizinhaca2DB - Texto original (Subáreas)

Período analisado	**V1**	**V2**
Últimos 3 Anos	42,83(+21,5)	42,23(+20,9)
Últimos 5 Anos	49,39(+23,1)	46,39(+20,1)
Produção Completa	-	-

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

(2013) para cálculo da classe resultante. Não foi computado resultados para a produção completa das Subáreas uma vez que o computador utilizado não tinha memória o suficiente para comportar os cálculos necessários.

Pode-se considerar os resultados obtidos nessa abordagem como o *baseline* para os resultados de análise de rede social. Observa-se que os resultados de V2 foram inferiores aos de V1, dando indícios de que a vizinhança de nível um é mais adequada que a nível dois.

Tabela 22 – Resultados do conjunto Tfidf+V1DB (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	91,45(+0,61)	91,88(+1,04)	93,48(+2,64)	93,16(+2,32)
	Traduzido	90,91(+0,07)	90,59(-0,25)	92,73(+1,89)	92,30(+1,46)
Área	Normal	83,70(+1,59)	67,40(-14,7)	85,90(+3,79)	87,22(+5,11)
	Traduzido	78,12(-3,99)	63,58(-18,5)	81,64(-0,47)	82,81(+0,7)
Subárea	Normal	45,48(+12,9)	31,35(-1,23)	38,70(+6,12)	51,69(+19,1)
	Traduzido	42,93(+10,3)	27,11(-5,47)	38,13(+5,55)	47,17(+14,5)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Adicionalmente, os resultados dos conjuntos Tfidf+V1DB e Tfidf+V2DB estão presentes nas tabelas 22 e 23. As acurácias foram superiores às obtidas no *baseline* para a Análise de Rede Social, o que indica que a abordagem adotada é adequada para o problema.

Outro ponto está relacionado a comparação entre usar vizinhança de nível um (V1) e usar vizinhança de nível dois (V2), cujos resultados dão indícios de que, igualmente ao *baseline*, a vizinhança de nível um (V1) é mais adequada para o problema do que a de nível dois, uma vez que retorna acurácias maiores.

Além disso, tal abordagem conseguiu acurácias superiores aos resultados encontrados nas tabelas 6, 7 e 12, presentes nas seções 6.2 e 6.3, demostrando evidências de que usar as

Tabela 23 – Resultados do conjunto Tfidf+V2DB (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	89,74(-1,1)	87,92(-2,92)	91,66(+0,82)	89,31(-1,53)
	Traduzido	86,43(-4,41)	87,07(-3,77)	89,10(-1,74)	86,43(-4,41)
Área	Normal	83,70(+1,59)	67,40(-14,7)	85,90(+3,79)	87,22(+5,11)
	Traduzido	68,28(-13,8)	59,32(-22,7)	75,62(-6,49)	76,35(-5,76)
Subárea	Normal	33,33(+0,75)	26,83(-5,75)	36,15(+3,57)	44,06(+11,4)
	Traduzido	34,74(+2,16)	29,66(-2,92)	35,31(+2,73)	47,74(+15,1)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

métricas de análise de rede social ajudam no desempenho dos algoritmos. Por exemplo, na classificação das áreas a acurácia foi de $81,792\%$, na abordagem original, para $87,225\%$, na abordagem que soma os valores de V1 e V2 ao TF-IDF.

Tabela 24 – Resultados do conjunto Language-TfidfV1DB (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	89,85(-0,99)	88,24(-2,6)	92,84(+2)	92,20(+1,36)
	Traduzido	90,38(-0,46)	88,78(-2,06)	93,80(+2,96)	90,49(-0,35)
Área	Normal	65,34(-16,7)	69,31(-12,8)	67,54(-14,5)	84,72(+2,61)
	Traduzido	65,49(-16,6)	71,21(-10,9)	66,81(-15,3)	83,26(+1,15)
Subárea	Normal	25,98(-6,6)	30,50(-2,08)	35,59(+3,01)	51,69(+19,1)
	Traduzido	26,27(-6,31)	29,09(-3,49)	34,74(+2,16)	53,10(+20,5)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Tabela 25 – Resultados do conjunto Language-TfidfV2DB (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	90,27(-0,57)	87,50(-3,34)	91,45(+0,61)	91,02(+0,18)
	Traduzido	88,24(-2,6)	88,03(-2,81)	93,05(+2,21)	92,20(+1,36)
Área	Normal	64,17(-17,9)	64,46(-17,6)	66,66(-15,4)	82,52(+0,41)
	Traduzido	65,49(-16,6)	71,21(-10,9)	66,81(-15,3)	83,26(+1,15)
Subárea	Normal	25,70(-6,88)	30,50(-2,08)	33,61(+1,03)	49,71(+17,1)
	Traduzido	24,85(-7,73)	29,37(-3,21)	32,20(-0,38)	45,76(+13,1)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Os conjuntos Language-TfidfV1DB e Language-TfidfV2DB têm seus resultados representados nas tabelas 24 e 25. Possuindo uma abordagem similar aos conjuntos

Tfidf+VXDB, com a diferença de que ao invés de somar as métricas de rede social a representação numérica por TF-IDF, esses conjuntos estudam a contribuição da rede social junto da proporção dos idiomas na produção científica de cada pesquisador.

Destaca-se que os resultados encontrados não foram superiores aos encontrados nos conjuntos Tfidf+VXDB, apesar de serem melhores do que a representação numérica por si só, cujos resultados estão presentes nas tabelas 6, 7 e 12.

Tabela 26 – Resultados do conjunto Wiki-V1-150 (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	84,83(-6,01)	78,10(-12,7)	89,60(-1,24)	91,68(+0,84)
	Traduzido	84,56(-6,28)	83,57(-7,27)	90,40(-0,44)	92,86(+2,02)
Área	Normal	63,28(-18,8)	63,63(-18,4)	69,30(-12,8)	80,16(-1,95)
	Traduzido	65,65(-16,4)	70,39(-11,7)	74,05(-8,06)	84,98(+2,87)
Subárea	Normal	30,93(-1,65)	40,96(+8,38)	37,70(+5,12)	53,44(+20,8)
	Traduzido	31,57(-1,01)	48,30(+15,7)	44,08(+11,5)	57,53(+24,9)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Tabela 27 – Resultados do conjunto Wiki-V2-150 (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	82,94(-7,9)	77,37(-13,4)	88,76(-2,08)	91,27(+0,43)
	Traduzido	83,29(-7,55)	82,39(-8,45)	89,69(-1,15)	92,49(+1,65)
Área	Normal	59,05(-23,0)	60,71(-21,4)	69,08(-13,1)	78,87(-3,24)
	Traduzido	63,35(-18,7)	66,65(-15,4)	73,85(-8,26)	84,55(+2,44)
Subárea	Normal	29,77(-2,81)	41,05(+8,47)	38,42(+5,84)	52,14(+19,5)
	Traduzido	31,70(-0,88)	46,30(+13,7)	43,86(+11,2)	56,87(+24,2)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Por fim, o último teste com as métricas de rede social envolveu a representação numérica por meio de tópicos, aliada aos valores de vizinhança, representada nos conjuntos Wiki-VX-Y. As tabelas 26 e 27 possuem as acurácias para as representações com 150 tópicos, da mesma forma que as tabelas 28 e 29 mostram os resultados com 300 tópicos.

No geral, as acurácias foram consideravelmente maiores do que a abordagem original (tabelas 13 e 14), com novamente a representação com 300 tópicos sendo melhor do que a com apenas 150 tópicos. Além disso, as acurácias foram tão boas quanto (no sentido

Tabela 28 – Resultados do conjunto Wiki-V1-300 (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
Grande Área	Normal	84,53(-6,31)	79,06(-11,7)	89,01(-1,83)	92,42(+1,58)
	Traduzido	84,48(-6,36)	84,57(-6,27)	90,44(-0,4)	93,37(+2,53)
Área	Normal	66,56(-15,5)	66,94(-15,1)	70,40(-11,7)	83,76(+1,65)
	Traduzido	66,94(-15,1)	72,64(-9,47)	74,47(-7,64)	86,81(+4,7)
Subárea	Normal	30,17(-2,41)	41,18(+8,6)	38,22(+5,64)	56,47(+23,8)
	Traduzido	32,04(-0,54)	50,37(+17,7)	43,84(+11,2)	61,05(+28,4)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

de serem valores muito próximos) as obtidas nos conjuntos Tfidf+VXDB, superando até mesmo o Language-TfidfVXDB, nos resultados com 300 tópicos.

Tabela 29 – Resultados do conjunto Wiki-V2-300 (toda produção científica) - Acurácia (%)

Nível	**Conjunto**	**J48**	**Árvore de Decisão**	**Random Forest**	**SVM**
Grande Área	Normal	83,82(-7,02)	79,10(-11,7)	88,24(-2,6)	92,06(+1,22)
	Traduzido	82,83(-8,01)	83,98(-6,86)	90,06(-0,78)	93,27(+2,43)
Área	Normal	62,97(-19,1)	66,62(-15,4)	69,67(-12,4)	82,74(+0,63)
	Traduzido	64,73(-17,3)	71,80(-10,3)	74,26(-7,85)	86,60(+4,49)
Subárea	Normal	29,85(-2,73)	42,14(+9,56)	38,49(+5,91)	55,83(+23,2)
	Traduzido	30,91(-1,67)	49,33(+16,7)	43,71(+11,1)	61,10(+28,5)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Uma vez que praticamente todos os testes envolvendo as métricas de rede social, especialmente a vizinhança de nível um (V1), tiveram acurácias superiores às abordagens originais e ao próprio *baseline* usando métricas de rede social, pode-se argumentar de que utilizar de tais métricas melhora o desempenho dos algoritmos testados. Contudo, para ter convicção de tal argumento, os melhores resultados de todas abordagens são comparados usando testes estatísticos na seção 6.8.

6.6 Classificação em dois níveis

Os resultados utilizando a classificação em dois níveis, explicada na seção 4.8, são discutidos a seguir. Apenas o problema das Grandes Áreas foi considerado, assim como apenas se avaliou dois algoritmos (aqueles que retornaram as melhores acurácias nos outros

testes): *Random Forest* e *Support Vector Machine*. Os conjuntos TfIdfDB, Wiki-150 e Wiki-300 foram os escolhidos para a criação dos subconjuntos (usados no treinamento e teste da classificação em dois níveis), por apresentarem bons resultados no geral e não possuírem outros atributos que possam influenciar no resultado (como métricas de rede social, por exemplo). Assim, os conjuntos resultantes HierarquicoDB, HierarquicoWiki-150 e HierarquicoWiki-300 foram criados e adotados na classificação em dois níveis.

Tabela 30 – Comparação entre os algoritmos com classificação em dois níveis (com TF-IDF) - Acurácia (%)

Período analisado	**Conjunto**	**Random Forest (Padrão)**	**Random Forest (em Hierarquia)**	**SVM (Padrão)**	**SVM (em Hierarquia)**
3 Anos	Normal	90,22(+2,15)	90,29(+2,22)	90,34(+2,27)	89,97(+1,9)
	Traduzido	88,12(+0,05)	87,46(-0,61)	89,17(+1,1)	88,53(+0,46)
	Enriquecido	82,49(-5,58)	81,33(-6,74)	79,50(-8,57)	80,43(-7,64)
5 Anos	Normal	89,77(+0,94)	89,29(+0,46)	90,36(+1,53)	90,27(+1,44)
	Traduzido	91,24(+2,41)	90,71(+1,88)	90,24(+1,41)	90,92(+2,09)
	Enriquecido	83,76(-5,07)	83,54(-5,29)	84,19(-4,64)	83,43(-5,4)
Completo	Normal	90,72(-0,12)	90,80(-0,04)	92,22(+1,38)	92,30(+1,46)
	Traduzido	91,72(+0,88)	90,99(+0,15)	91,52(+0,68)	91,75(+0,91)
	Enriquecido	87,69(-3,15)	87,28(-3,56)	88,89(-1,95)	87,92(-2,92)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

A tabela 30 mostra o resultado dos classificadores hierárquicos usando o conjunto HierarquicoDB, enquanto que a tabela 31 apresenta os resultados para os conjuntos HierarquicoWiki-150 e HierarquicoWiki-300 (usando toda produção científica).

Um ponto considerado e que foi motivação para o desenvolvimento do classificador hierárquico (em dois níveis), é que nesse contexto o enriquecimento de texto poderia retornar acurácias altas, pois como ele adiciona ambiguidade no texto (no sentido de reforçar classes muito similares), acreditava-se que ao classificar primeiro uma instância em humanas/não-humanas o enriquecimento iria ajudar a separar bem essas duas classes e, subsequentemente, os classificadores específicos conseguiriam inferir adequadamente a classe resultante.

Contudo, ao analisar os resultados na tabela 30, nota-se tal fato não ocorreu, uma vez que os resultados encontrados foram similares a não adotar a classificação em dois níveis (em questão de acurácia). Portanto, o texto enriquecido continua não melhorando a acurácia dos algoritmos.

Tabela 31 – Comparação entre os algoritmos com classificação em dois níveis (por tópicos) - Acurácia (%)

Número de tópicos	**Conjunto**	**Random Forest (Padrão)**	**Random Forest (em Hierarquia)**	**SVM (Padrão)**	**SVM (em Hierarquia)**
150 tópicos	Normal	84,60(-6,24)	84,94(-5,9)	87,57(-3,27)	87,56(-3,28)
	Traduzido	86,84(-4)	87,14(-3,7)	90,08(-0,76)	89,94(-0,9)
300 tópicos	Normal	85,40(-5,44)	85,63(-5,21)	89,81(-1,03)	90,06(-0,78)
	Traduzido	87,46(-3,38)	87,58(-3,26)	91,89(+1,05)	91,70(+0,86)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Observando as tabelas 30 e 31, outro ponto interessante é perceptível: no geral a abordagem não conseguiu melhorar os resultados dos algoritmos (Random Forest e SVM), uma vez que a variação das acurácias foi menor que 1%, tanto em melhoria quanto em piora.

6.7 Uso da informação de hierarquia

Nesta seção, os resultados utilizando do atributo adicional "classe superior" são apresentados e sua potencial influência no desempenho dos algoritmos será discutida. Os conjuntos TfidfDB (usando toda produção científica), Wiki-150 e Wiki-300 tiveram seus resultados computados com esse atributo adicionado aos mesmos, criando os conjuntos HierarquiaClasseDB, HierarquiaClasseWiki-150 e HierarquiaClasseWiki-300, respectivamente. Os algoritmos J48, *Naive Bayes*, *Random Forest* e SVM foram os escolhidos para a avaliação.

Tabela 32 – Resultados da classificação usando hierarquia de classes para as Áreas - Acurácia (%)

Abordagem	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
TF-IDF	Normal	78,85(-3,26)	65,05(-17,0)	74,44(-7,67)	86,05(+3,94)
	Traduzido	79,73(-2,38)	67,84(-14,2)	75,03(-7,08)	86,93(+4,82)
150 tópicos	Normal	66,71(-15,4)	53,00(-29,1)	76,22(-5,89)	75,61(-6,5)
	Traduzido	75,92(-6,19)	67,87(-14,2)	82,49(+0,38)	88,78(+6,67)
300 tópicos	Normal	72,96(-9,15)	58,67(-23,4)	75,48(-6,63)	84,80(+2,69)
	Traduzido	77,84(-4,27)	70,24(-11,8)	80,59(-1,52)	91,16(+9,05)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

A tabela 32 mostra as acurácia para o problema de inferência das Áreas, assim como a tabela 33 mostra os resultados para as Subáreas. Comparando com os resultados vistos nas seções anteriores, percebe-se um aumento considerável na acurácia no geral.

Por exemplo, para as Áreas o melhor resultado encontrado até então no conjunto TfidfDB tinha sido de $81,79\%$ de acurácia, $82,24\%$ de acurácia para o conjunto Wiki-150 e $85,80\%$ de acurácia no conjunto Wiki-300. Tais acurácias comparadas com as presentes na tabela 32, mostram uma melhoria absoluta na taxa de acerto de $5,14\%$, $6,54\%$ e $5,36\%$, respectivamente e um aumento percentual de $6,28\%$, $7,95\%$ e $6,24\%$, nos conjuntos com o atributo "classe superior".

Tabela 33 – Resultados da classificação usando hierarquia de classes para as Subáreas - Acurácia (%)

Abordagem	**Conjunto**	**J48**	**Naive Bayes**	**Random Forest**	**SVM**
TF-IDF	Normal	47,74(+15,1)	29,94(-2,64)	38,98(+6,4)	59,88(+27,3)
	Traduzido	48,02(+15,4)	28,81(-3,77)	38,98(+6,4)	60,45(+27,8)
150 tópicos	Normal	55,14(+22,5)	34,72(+2,14)	63,27(+30,6)	62,53(+29,9)
	Traduzido	58,57(+25,9)	47,41(+14,8)	67,53(+34,9)	69,48(+36,9)
300 tópicos	Normal	55,81(+23,2)	37,09(+4,51)	55,81(+23,2)	66,79(+34,2)
	Traduzido	57,78(+25,2)	48,03(+15,4)	63,86(+31,2)	73,79(+41,2)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Igualmente, os melhores resultados para os conjuntos TfidfDB, Wiki-150 e Wiki-300 no problema das Subáreas foram, respectivamente, $52,41\%$, $53,44\%$ e $59,48\%$ de acurácia até então. Comparando-os com os presentes na tabela 33, identifica-se uma melhoria absoluta de $8,04\%$, $16,04\%$ e $14,31\%$ nas melhores acurácias encontradas, respectivamente, além de uma melhoria percentual de $15,34\%$, $30,01\%$ e $24,05\%$ com as mesmas acurácias.

Os resultados apresentados fornecem indícios de que utilizar a informação da classe acima na estrutura das classes melhora consideravelmente o desempenho dos algoritmos. Uma das hipóteses para explicar isso se dá no fato de que esse tipo de informação diminui o tamanho do problema que os algoritmos tem de lidar, pois eles podem dar mais ênfase as classes mais relacionadas a mesma (por exemplo, caso a classe do nível acima seja Ciências Exatas e da Terra, os algoritmos poderão dar mais ênfase na avaliação das áreas relacionadas a ela). Assim, como o problema é menor a taxa de acerto deles pode aumentar consideravelmente, o que explicaria as acurácias obtidas até então.

6.8 Testes Estatísticos

Nesta seção são apresentados testes estatísticos sobre dos resultados das melhores abordagens (ou seja, aquelas que retornaram as melhores acurácias) encontradas até então. Como método para avaliar se existe uma diferença estatisticamente relevante sobre as acurácias de duas abordagens distintas, ou seja determinar se uma é possivelmente melhor do que a outra, o teste não paramétrico de Friedman (também chamado de $p - value$) será aplicado sobre os resultados obtidos para determinar quais abordagens são as melhores.

Dada uma hipótese nula H_o, o $p - value$ é a probabilidade da diferença entre as médias amostrais entre dois grupos ser igual ou maior do que o extremo do valor observado (WASSERSTEIN; LAZAR, 2016) sobre a hipótese nula. No caso do teste, busca-se encontrar um $p - value$ pequeno sobre a hipótese nula, de forma que a hipótese nula seja rejeitada com um certo grau de confiança.

Para os testes que foram empregados, duas hipóteses são propostas: uma hipótese nula (H_o) e uma hipótese alternativa (H_a). O $p - value$ será computado a partir da hipótese nula, que representará a hipótese de que as duas abordagens testadas em cada teste possuam a mesma acurácia média (em porcentagem), ou seja, dadas duas médias μ_1 e μ_2, a hipótese nula é que $\mu_1 - \mu_2 = 0$. A hipótese alternativa em todos os testes será de que as acurácias médias são diferentes, ou seja, uma das abordagens possui acurácia maior do que a outra ($\mu_1 - \mu_2 > 0$).

O $p - value$ é computado a partir das acurácias, assim é necessário saber o desvio padrão das abordagens para se obter uma variável chamada de Z_{obs}, dada pela equação 8. Nesta, $\overline{y_1}$ e $\overline{y_2}$ são as médias amostrais que são avaliadas, s_1 e s_2 seus respectivos desvios padrões e n_1 e n_2 o número de elementos usados para calcular tais médias. Computa-se o valor $P(Z > Z_{obs})$ para encontrar o $p - value$, sendo Z a probabilidade do intervalo de confiança testado.

$$Z_{obs} = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{8}$$

O $p - value$ calculado, no caso desse trabalho, terá um intervalo de confiança de 95%, ou seja, o $p - value$ deverá ser menor que $0,05$ para que a hipótese nula seja rejeitada. Outra forma também de rejeitar a hipótese nula é olhando para Z_{obs}, pois caso Z_{obs} seja maior que $Z_{0,05} = 1,645$, isso implica que $p - value < 0,05$, logo a hipótese nula é rejeitada.

Para os testes dessa seção, os dois valores são considerados na hora de avaliar se a hipótese nula deve ser rejeitada ou não.

Dessa forma, os conjuntos que retornaram melhores acurácias foram selecionados para serem avaliados utilizando 10-*fold cross validation* novamente, só que dessa vez computando além da acurácia média, qual o desvio padrão das acurácias obtidas. As tabelas 34, 35 e 36 mostram as acurácias médias e seus respectivos desvios padrões (em porcentagem) nos três problemas tratados nessa pesquisa: Grandes Áreas, Áreas e Subáreas.

Tabela 34 – Acurácias e desvios padrões utilizados nos testes estatísticos (Grandes Áreas)

Abordagem	**Conjunto**	**Acurácia(%)**	**Desvio Padrão(%)**
Baseline	Normal	90,84	0,9
	Traduzido	90,77	0,8
TfidfDB	Normal	92,22	2,8
	Traduzido	91,52	2,86
Wiki-150	Normal	87,57	1,15
	Traduzido	90,08	0,88
Wiki-300	Normal	89,81	1,08
	Traduzido	91,89	0,78
Wiki-V1-150	Normal	91,74	0,85
	Traduzido	92,83	0,87
Wiki-V2-150	Normal	91,25	0,91
	Traduzido	92,41	0,8
Wiki-V1-300	Normal	92,39	0,85
	Traduzido	93,45	0,86
Wiki-V2-300	Normal	92,03	0,82
	Traduzido	93,25	0,93
Language-TfidfV1	Normal	92,2	2,57
	Traduzido	93,46	2,53
Language-TfidfV2	Normal	91,38	2,87
	Traduzido	92,04	2,67
HierarquicoDB*	Normal	92,30	2,77
	Traduzido	91,75	2,31
HierarquicoWiki-150*	Normal	87,56	0,87
	Traduzido	89,94	0,76
HierarquicoWiki-300*	Normal	90,06	1,40
	Traduzido	91,70	0,65

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Nota-se que apenas foram utilizadas as melhores abordagens (aquelas com maior acurácia) encontradas até então, ou seja, toda a produção científica completa dos pesquisadores foi adotada em todos os testes, além do classificador *Support Vector Machine* para

a maioria dos conjuntos, com exceção do *Baseline* que usa a abordagem de Miyata, Kano e Digiampietri (2013) para fazer a classificação.

Tabela 35 – Acurácias e desvios padrões utilizados nos testes estatísticos (Áreas)

Abordagem	**Conjunto**	**Acurácia**(%)	**Desvio Padrão**(%)
Baseline	Normal	82,11	1,3
	Traduzido	80,58	1,2
TfidfDB	Normal	81,75	3,32
	Traduzido	80	3,16
Wiki-150	Normal	65,99	1,5
	Traduzido	82,26	1,34
Wiki-300	Normal	76,64	1,45
	Traduzido	85,88	1,39
Wiki-V1-150	Normal	80,27	1,37
	Traduzido	84,92	1,11
Wiki-V2-150	Normal	78,75	1,36
	Traduzido	84,59	1,12
Wiki-V1-300	Normal	83,75	1,33
	Traduzido	86,71	1,22
Wiki-V2-300	Normal	82,78	1,36
	Traduzido	86,69	1,13
Language-TfidfV1	Normal	84,76	3,81
	Traduzido	83,64	3,6
Language-TfidfV2	Normal	82,98	3,63
	Traduzido	82,5	3,85
HierarquiaClasseDB	Normal	86,3	3,15
	Traduzido	86,77	2,74
HierarquiaClasseWiki-150	Normal	75,45	1,15
	Traduzido	88,76	1,11
HierarquiaClasseWiki-300	Normal	84,68	1,14
	Traduzido	91,36	1,16

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Algumas acurácias estão diferentes das apresentadas nas seções anteriores, uma vez que o 10-*fold cross validation* foi aplicado novamente sobre os conjuntos, pois anteriormente não estava sendo computado o desvio padrão da acurácia média dos testes. Uma vez tendo as acurácias e os desvios padrões, basta escolher quais abordagens são comparadas com quais e computar o $p - value$ da hipótese nula sugerida, de forma a delinear se uma das duas abordagens comparadas tem uma acurácia estatisticamente melhor do que a outra, ou não.

Tabela 36 – Acurácias e desvios padrões utilizados nos testes estatísticos (Subáreas)

Abordagem	**Conjunto**	**Acurácia(%)**	**Desvio Padrão(%)**
Baseline	Normal	32,58	3,2
	Traduzido	31,47	2,4
TfidfDB	Normal	51,62	6,59
	Traduzido	50,26	7,28
Wiki-150	Normal	39,91	1,47
	Traduzido	53,62	1,73
Wiki-300	Normal	47	1,48
	Traduzido	59,77	1,57
Wiki-V1-150	Normal	53,25	1,77
	Traduzido	57,57	1,74
Wiki-V2-150	Normal	51,78	1,84
	Traduzido	56,87	1,81
Wiki-V1-300	Normal	56,64	1,8
	Traduzido	61,13	1,66
Wiki-V2-300	Normal	55,93	1,93
	Traduzido	61,02	1,81
Language-TfidfV1	Normal	51,93	6,4
	Traduzido	51,63	6,35
Language-TfidfV2	Normal	48,56	7,05
	Traduzido	46,33	6,59
HierarquiaClasseDB	Normal	60,77	6,22
	Traduzido	59,76	7,27
HierarquiaClasseWiki-150	Normal	62,69	1,14
	Traduzido	69,7	1,42
HierarquiaClasseWiki-300	Normal	66,85	1,51
	Traduzido	73,87	1,52

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Contudo, para efeito de organização, os testes feitos foram divididos em três grupos, referentes a cada um dos níveis da taxonomia de áreas de atuação do Lattes. Assim, um dos grupos irá avaliar quais abordagens são melhores para as Grandes Áreas, outro avaliará as abordagens na tarefa de inferência das Áreas e por último um grupo de testes avaliará as abordagens nas Subáreas. Cada um desses grupos são discutidos nas subseções 6.8.1, 6.8.2 e 6.8.3

6.8.1 Grandes Áreas

Os testes estatísticos para as Grandes Áreas tomam as acurácias e os desvios padrões da tabela 34 para calcular o $p - value$ dos testes. Os testes feitos e seus objetivos são os seguintes (hipóteses alternativas):

1. TfidfDB (Normal) > Baseline (Normal): Avaliar se a representação numérica por TF-IDF é melhor do que o *Baseline*.
2. Wiki-300 (Traduzido) > Baseline (Normal): Avaliar se a representação numérica usando 300 tópicos é melhor do que o *Baseline*.
3. HierarquicoDB (Normal) > Baseline (Normal): Avaliar se a classificação em dois níveis usando a representação numérica por TF-IDF é melhor do que o *Baseline*.
4. Wiki-300 (Traduzido) > Wiki-150 (Traduzido): Avaliar se usar 300 tópicos na representação numérica por tópicos é melhor do que somente usar 150.
5. Wiki-150 (Traduzido) > Wiki-150 (Normal): Avaliar se traduzir o texto melhora o desempenho da representação numérica por tópicos (150).
6. Wiki-300 (Traduzido) > Wiki-300 (Normal): Avaliar se traduzir o texto melhora o desempenho da representação numérica por tópicos (300).
7. TfidfDB (Normal) > Wiki-300 (Traduzido): Avaliar se a representação numérica por TF-IDF é melhor representação numérica por tópicos (300).
8. Wiki-V1-150 (Traduzido) > Wiki-V2-150 (Traduzido): Avaliar se a métrica de rede social de uma vizinhança (V1) é melhor do que a duas vizinhanças (V2) para a representação numérica por tópicos (150).
9. Wiki-V1-300 (Traduzido) > Wiki-V2-300 (Traduzido): Avaliar se a métrica de rede social de uma vizinhança (V1) é melhor do que a duas vizinhanças (V2) para a representação numérica por tópicos (300).
10. Language-TfidfV1 (Traduzido) > Wiki-V1-300 (Traduzido): Avaliar se a representação numérica por TF-IDF é do que melhor representação numérica por tópicos (300), ambas usando análise de rede social.
11. Language-TfidfV1 (Traduzido) > TfidfDB (Normal): Avaliar se a representação numérica por TF-IDF usando análise de rede social é melhor do que do mesma sem usar.

12. Wiki-V1-300 (Traduzido) > Wiki-300 (Traduzido): Avaliar se a representação numérica por tópicos (300) usando análise de rede social (V1) é melhor do que a mesma sem usar.
13. HierarquicoDB (Normal) > TfidfDB (Normal): Avaliar se classificação em dois níveis, usando a representação numérica por TF-IDF, possui resultados melhores do que a mesma abordagem sem classificação em dois níveis.
14. Wiki-150 (Traduzido) > HierarquicoWiki-150 (Traduzido) : Avaliar se a classificação em dois níveis, usando a representação numérica por 150 tópicos, é pior que a mesma abordagem sem classificação em dois níveis.
15. HierarquicoWiki-300 (Traduzido) > Wiki-300 (Traduzido): Avaliar se a classificação em dois níveis, usando a representsação numérica por 300 tópicos, é melhor do que a mesma abordagem sem classificação em dois níveis.
16. HierarquicoDB (Normal) > HierarquicoWiki-300 (Traduzido): Avaliar se a classificação em dois níveis, usando a representação numérica por TF-IDF, é melhor do que a classificação em dois níveis usando a representação numérica por 300 tópicos.
17. HierarquicoDB (Normal) > Wiki-V1-300 (Traduzido): Avaliar se a classificação em dois níveis, usando a representação numérica por TF-IDF, é melhor do que a representação numérica por 300 tópicos usando análise de rede social de vizinhança um (V1).
18. Wiki-V1-300 (Traduzido) > HierarquicoWiki-300 (Traduzido): Avaliar se usar usar informação de rede social (V1) é melhor do que usar classificação em dois níveis.

Cada um dos testes feitos tentou elucidar qual dentre as abordagens comparadas possuem os melhores resultados no geral. Nota-se que as hipóteses propostas são as hipóteses alternativas e que as hipóteses nulas referentes a cada um dos testes é aquela que considera que as duas abordagens são iguais, ou seja, que não existe diferença estatística substancial nas suas acurácias médias para que alguma conclusão seja tirada.

A tabela 37, mostra os valores de Z_{obs} e $p - value$ para os testes feitos a partir dos resultados das Grandes Áreas. Nota-se que para uma hipótese nula ser negada, isso significa que $Z_{obs} > 1,645$ e $p - value < 0,05$, caso o contrário os resultados são inconclusivos, não sendo possível tirar qualquer tipo de conclusão.

Caso a hipótese nula seja negada, a hipótese alternativa apresentada anteriormente é verdadeira. Por exemplo, os testes 1, 2 e 3 avaliam algumas das propostas alternativas

Tabela 37 – Resultados dos testes estatísticos - Grandes Áreas

Teste	Z_{obs}	$p - value$	**Conclusão**
1	1,4838	0,068933	Resultados inconclusivos
2	2,788	0,0026519	Hipótese nula foi negada
3	1.5852	0.056462	Resultados inconclusivos
4	4,8674	5,6532e-07	Hipótese nula foi negada
5	5,4813	2,1109e-08	Hipótese nula foi negada
6	4,9373	3,9608e-07	Hipótese nula foi negada
7	0,35903	0,35979	Resultados inconclusivos
8	1,1237	0,13056	Resultados inconclusivos
9	0,4993	0,30878	Resultados inconclusivos
10	0,011834	0,49528	Resultados inconclusivos
11	1,0391	0,14938	Resultados inconclusivos
12	4,2489	1,074e-05	Hipótese nula foi negada
13	0,064231	0,47439	Resultados inconclusivos
14	0,38075	0,35169	Resultados inconclusivos
15	0,59176	0,72299	Resultados inconclusivos
16	0,66686	0,25243	Resultados inconclusivos
17	1,2538	0,10495	Resultados inconclusivos
18	5,1335	1,4218e-07	Hipótese nula foi negada

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

deste trabalho, comparando-as com o *baseline*. Nesse caso, os resultados mostram que não se pode afirmar se a representação numérica por TF-IDF é melhor do que o baseline, pois os resultados foram inconclusivos. Contudo, pode-se fazer essa afirmação para a representação numérica usando 300 tópicos e para a classificação em dois níveis usando a representação numérica por TF-IDF.

Da mesma forma, o teste 4 avalia se conjunto o Wiki-300 (Traduzido) é melhor (consegue produzir resultados mais acurados) do que o conjunto Wiki-150 (Traduzido) em questão de acurácia média. A hipótese nula nesse caso afirma que Wiki-300 (Traduzido) é igual a Wiki-150 (Traduzido). Todavia, como a mesma foi negada, isso implica dizer que a hipótese alternativa está correta, já que a acurácia da SVM no conjunto Wiki-300 (Traduzido) é de fato maior que a observada no conjunto Wiki-150 (Traduzido), ou seja, Wiki-300 (Traduzido) > Wiki-150 (Traduzido).

No geral, os resultados apresentados na tabela 37 dão indícios de algumas possíveis conclusões sobre as abordagens testadas. O quadro 5 mostra um resumo de quais hipóteses alternativas estão corretas. No que diz respeito aos testes feitos sobre o problema das Grandes Áreas, usar um número maior de tópicos na representação numérica baseada em tópicos melhora o desempenho dos algoritmos (teste 4), bem como usar o texto traduzido

Quadro 5 – Quadro de Hipóteses Alternativas - Grandes Áreas

Teste	**Hipóteses Alternativas**
2	Wiki-300 (Traduzido) > Baseline (Normal)
4	Wiki-300 (Traduzido) > Wiki-150 (Traduzido)
5	Wiki-150 (Traduzido) > Wiki-150 (Normal)
6	Wiki-150 (Traduzido) > Wiki-150 (Normal)
12	Wiki-V1-300 (Traduzido) > Wiki-300 (Traduzido)
18	Wiki-V1-300 (Traduzido) > HierarquicoWiki-300 (Traduzido)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

dos títulos da produção científica melhora também o desempenho dos algoritmos (testes 5 e 6).

Além disso, não é possível se tirar nenhuma conclusão sobre qual das métricas de rede social retorna o melhor desempenho (testes 8 e 9), uma vez que os resultados obtidos nos testes estatísticos foram inconclusivos. Da mesma forma, não é possível se concluir qual das abordagens numéricas tem o melhor desempenho com a métrica V1 (teste 10), bem como não se pode afirmar que o uso da métrica de rede social V1 melhora o resultado da representação numérica por TF-IDF (teste 11).

Contudo, o teste 12 mostra que usar a métrica de rede social V1 com a representação numérica por tópicos melhora o desempenho dos algoritmos. Usar a classificação em dois níveis não melhora o desempenho dos algoritmos (testes 13, 14 e 15), e não se pode fazer nenhuma conclusão sobre qual delas é a melhor (teste 16). Além disso, o uso da classificação em dois níveis (por TF-IDF) não possui resultados melhores do que usar a métrica de rede social V1 (teste 17), porém a métrica de rede social V1 é melhor do que classificação em dois níveis quando esta usa 300 tópicos (teste 18).

6.8.2 Áreas

As hipóteses alternativas avaliadas para o problema das Áreas, da taxonomia de áreas de atuação do Lattes, são as seguintes:

1. Baseline (Normal) > TfidfDB (Normal): Avaliar se o *Baseline* é melhor do que representação numérica por TF-IDF.
2. Wiki-300 (Traduzido) > Baseline (Normal): Avaliar se a representação numérica usando 300 tópicos é melhor do que o *Baseline*.

3. HierarquiaClasseDB (Traduzido) > Baseline (Normal): Avaliar se adotar informação de hierarquia das classes com a representação numérica por TF-IDF possui resultados melhores do que o *Baseline*.
4. Wiki-300 (Traduzido) > Wiki-150 (Traduzido): Avaliar se usar 300 tópicos na representação numérica por tópicos é melhor do que somente usar 150.
5. Wiki-150 (Traduzido) > Wiki-150 (Normal): Avaliar se traduzir o texto melhora o desempenho da representação numérica por tópicos (150).
6. Wiki-300 (Traduzido) > Wiki-300 (Normal): Avaliar se traduzir o texto melhora o desempenho da representação numérica por tópicos (300).
7. Wiki-300 (Traduzido) > TfidfDB (Normal): Avaliar se a representação numérica por tópicos (300) é melhor do que representação numérica por TF-IDF.
8. Wiki-V1-150 (Traduzido) > Wiki-V2-150 (Traduzido): Avaliar se a métrica de rede social de uma vizinhança (V1) é melhor do que a duas vizinhanças (V2) para a representação numérica por tópicos (150).
9. Wiki-V1-300 (Traduzido) > Wiki-V2-300 (Traduzido): Avaliar se a métrica de rede social de uma vizinhança (V1) é melhor do que a duas vizinhanças (V2) para a representação numérica por tópicos (150).
10. Wiki-V1-300 (Traduzido) > Language-TfidfV1 (Traduzido) : Avaliar se a representação numérica por tópicos (300) usando análise de rede social é melhor do que a representação numérica por TF-IDF.
11. Language-TfidfV1 (Traduzido) > TfidfDB (Normal): Avaliar se a representação numérica por TF-IDF usando análise de rede social é melhor do que mesma sem usar.
12. Wiki-V1-300 (Traduzido) > Wiki-300 (Traduzido): Avaliar se a representação numérica por tópicos (300) usando análise de rede social é melhor do que a mesma sem usar.
13. HierarquiaClasseDB (Traduzido) > TfidfDB (Normal): Avaliar se adotar informação de hierarquia das classes com a representação numérica por TF-IDF possui resultados melhores do que a mesma abordagem sem usar essa informação.
14. HierarquiaClasseWiki-150 (Traduzido) > Wiki-150 (Traduzido): Avaliar se adotar informação de hierarquia das classes com a representação numérica por 150 tópicos é melhor do que a mesma abordagem sem usar essa informação.

15. HierarquiaClasseWiki-300 (Traduzido) > Wiki-300 (Traduzido): Avaliar se adotar informação de hierarquia das classes com a representação numérica por 300 tópicos, é melhor do que a mesma abordagem sem usar essa informação.
16. HierarquiaClasseWiki-300 (Traduzido) > HierarquiaClasseDB (Traduzido): Avaliar se adotar informação de hierarquia das classes com a representação numérica por 300 tópicos é melhor do que a representação numérica por TF-IDF que também adota informação de hierarquia das classes.
17. HierarquiaClasseDB (Traduzido) > Wiki-V1-300 (Traduzido): Avaliar se adotar informação de hierarquia das classes com a representação numérica por TF-IDF é melhor do que a representação numérica por 300 tópicos usando análise de rede social de vizinhança um (V1).
18. HierarquiaClasseWiki-300 (Traduzido) > HierarquiaClasseWiki-150 (Traduzido): Avaliar se o número de tópicos influi na acurácia da classificação que usa informação de hierarquia das classes.

Tabela 38 – Resultados dos testes estatísticos - Áreas

Teste	Z_{obs}	$p - value$	**Conclusão**
1	0,31929	0,37475	Resultados inconclusivos
2	6,2641	1,8745e-10	Hipótese nula foi negada
3	4,859	5,8984e-07	Hipótese nula foi negada
4	5,9291	1,5231e-09	Hipótese nula foi negada
5	25,5797	0	Hipótese nula foi negada
6	14,5469	0	Hipótese nula foi negada
7	3,6286	0,00014248	Hipótese nula foi negada
8	0,66179	0,25405	Resultados inconclusivos
9	0,25405	0,48483	Resultados inconclusivos
10	1,5414	0,061611	Resultados inconclusivos
11	1,8835	0,029815	Hipótese nula foi negada
12	1,4192	0,077925	Resultados inconclusivos
13	3,6878	0,00011311	Hipótese nula foi negada
14	11,8129	0	Hipótese nula foi negada
15	9,5719	0	Hipótese nula foi negada
16	4,8782	5,352e-07	Hipótese nula foi negada
17	8,7348	0	Hipótese nula foi negada
18	5,121	1,5194e-07	Hipótese nula foi negada

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Os resultados dos testes estatísticos, para as Áreas, estão presentes na tabela 38. Assim como para as Grandes Áreas, as hipóteses nulas são negadas quando $Z_{obs} > 1,645$ e

$p - value < 0,05$. O quadro 6 mostra em resumo as hipóteses alternativas corretas para os testes feitos nas Áreas.

Quadro 6 – Quadro de hipóteses alternativas corretas - Áreas

Teste	**Hipóteses Alternativas**
2	Representação numérica por 300 tópicos é melhor do que o Baseline
3	Usar hierarquia de classes é melhor do que o Baseline
4	Usar 300 tópicos é melhor do que usar 150 tópicos na representação por tópicos
5	Texto traduzido é melhor do que ele sem traduzir na representação por 150 tópicos
6	Texto traduzido é melhor do que ele sem traduzir na representação por 300 tópicos
7	Representação numérica por 300 tópicos é melhor do que a por TF-IDF
11	Usar TF-IDF com a métrica de social V1 é melhor do que o TF-IDF
13	Usar informação de hierarquia (TF-IDF) é melhor do que só usar o TF-IDF
14	Usar informação de hierarquia (150 tópicos) é melhor do que só usar 150 tópicos
15	Usar informação de hierarquia (300 tópicos) é melhor do que só usar 300 tópicos
16	Usar informação de hierarquia (300 tópicos) é melhor do que só usar informação de hierarquia (TF-IDF)
17	Usar informação de hierarquia (TF-IDF) é melhor do que usar a métrica de rede social V1
18	Usar informação de hierarquia (300 tópicos) é melhor do que usar informação de hierarquia (150 tópicos)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Novamente, a representação numérica por TF-IDF comparada com o *baseline* teve resultados inconclusivos (teste 1), enquanto que as representações numéricas usando 300 tópicos (teste 2) e o uso de informação de hierarquia (teste 3) são melhores do que o *baseline*. Além disso, há indícios que um número maior de tópicos na representação numérica por tópicos melhora o desempenho dos algoritmos (teste 4), assim como traduzir o texto melhora o desempenho dos algoritmos nessa representação (testes 5 e 6). Diferentemente do visto nas Grandes Áreas, para as Áreas, a representação numérica baseada em tópicos (300 tópicos) é melhor do que a representação por TF-IDF (teste 7).

Segundo o teste, não é possível se tirar alguma conclusão sobre qual das métricas de rede social retorna o melhor desempenho (testes 8 e 9), além de não ser possível se concluir qual das abordagens numéricas tem o melhor desempenho com a métrica V1 (teste 10). Existem indícios também, que o uso da métrica de rede social V1 melhora o resultado da representação numérica por TF-IDF (teste 11). Contudo, como os resultados do teste 12

foram inconclusivos, não é possível fazer a mesma afirmação vale para a representação numérica por tópicos.

Os testes 13, 14 e 15 mostram que usar a informação de hierarquia das classes melhora o desempenho dos algoritmos, sendo que a melhor representação dentre elas é a que adota a representação numérica com 300 tópicos (teste 16). Ademais, o uso da informação de hierarquia das classes possui resultados melhores do que usar a métrica de rede social V1 (teste 17).

6.8.3 Subáreas

As hipóteses alternativas avaliadas para o problema das Subáreas são as seguintes:

1. TfidfDB (Normal) > Baseline (Normal): Avaliar se a representação numérica por TF-IDF é melhor do que o *Baseline*.
2. Wiki-300 (Traduzido) > Baseline (Normal): Avaliar se a representação numérica usando 300 tópicos é melhor do que o *Baseline*.
3. HierarquiaClasseDB (Traduzido) > Baseline (Normal): Avaliar se adotar informação de hierarquia das classes com a representação numérica por TF-IDF possui resultados melhores do que o *Baseline*.
4. Wiki-300 (Traduzido) > Wiki-150 (Traduzido): Avaliar se usar 300 tópicos na representação numérica por tópicos é melhor do que somente usar 150.
5. Wiki-150 (Traduzido) > Wiki-150 (Normal): Avaliar se traduzir o texto melhora o desempenho da representação numérica por tópicos (150).
6. Wiki-300 (Traduzido) > Wiki-300 (Normal): Avaliar se traduzir o texto melhora o desempenho da representação numérica por tópicos (300).
7. Wiki-300 (Traduzido) > TfidfDB (Normal): Avaliar se a representação numérica por tópicos (300) é melhor do que representação numérica por TF-IDF.
8. Wiki-V1-150 (Traduzido) > Wiki-V2-150 (Traduzido): Avaliar se a métrica de rede social de uma vizinhança (V1) é melhor do que a duas vizinhanças (V2) para a representação numérica por tópicos (150).
9. Wiki-V1-300 (Traduzido) > Wiki-V2-300 (Traduzido): Avaliar se a métrica de rede social de uma vizinhança (V1) é melhor do que a duas vizinhanças (V2) para a representação numérica por tópicos (300).

10. Wiki-V1-300 (Traduzido) > Language-TfidfV1 (Traduzido): Avaliar se a representação numérica por tópicos (300) usando análise de rede social é melhor do que a representação numérica por TF-IDF.
11. Language-TfidfV1 (Traduzido) > TfidfDB (Normal): Avaliar se a representação numérica por TF-IDF usando análise de rede social é melhor do que a mesma sem usar.
12. Wiki-V1-300 (Traduzido) > Wiki-300 (Traduzido): Avaliar se a representação numérica por tópicos (300) usando análise de rede social é melhor do que a mesma sem usar.
13. HierarquiaClasseDB (Normal) > TfidfDB (Normal): Avaliar se adotar informação de hierarquia das classes com a representação numérica por TF-IDF possui resultados melhores do que a mesma abordagem sem usar essa informação.
14. HierarquiaClasseWiki-150 (Traduzido) > Wiki-150 (Traduzido): Avaliar se adotar informação de hierarquia das classes com a representação numérica por 150 tópicos é melhor do que a mesma abordagem sem usar essa informação.
15. HierarquiaClasseWiki-300 (Traduzido) > Wiki-300 (Traduzido): Avaliar se adotar informação de hierarquia das classes com a representação numérica por 300 tópicos é melhor do que a mesma abordagem sem usar essa informação.
16. HierarquiaClasseWiki-300 (Traduzido) > HierarquiaClasseDB (Normal): Avaliar se adotar informação de hierarquia das classes com a representação numérica por 300 tópicos é melhor do que a representação numérica por TF-IDF que também adota informação de hierarquia das classes.
17. HierarquiaClasseDB (Traduzido) > Wiki-V1-300 (Traduzido): Avaliar se adotar informação de hierarquia das classes com a representação numérica por TF-IDF, é melhor do que a representação numérica por 300 tópicos usando análise de rede social de vizinhança um (V1).
18. HierarquiaClasseWiki-300 (Traduzido) > HierarquiaClasseWiki-150 (Traduzido): Avaliar se o número de tópicos influi na acurácia da classificação que usa informação de hierarquia das classes.

A tabela 38 mostra os resultados dos testes estatísticos para as Subáreas. As conclusões que podem ser tiradas são muito parecidas com as observadas para as Áreas,

uma vez que os testes feitos foram os mesmos. O quadro 7 mostra em resumo as hipóteses alternativas corretas para os testes feitos nas Subáreas.

Tabela 39 – Resultados dos testes estatísticos - Subáreas

Teste	Z_{obs}	$p - value$	**Conclusão**
1	8,2188	1,1102e-16	Hipótese nula foi negada
2	24,1226	0	Hipótese nula foi negada
3	12,7443	0	Hipótese nula foi negada
4	8,3247	0	Hipótese nula foi negada
5	19,0974	0	Hipótese nula foi negada
6	18,7162	0	Hipótese nula foi negada
7	3,8044	7,1079e-05	Hipótese nula foi negada
8	0,88166	0,18898	Resultados inconclusivos
9	0,14164	0,44368	Resultados inconclusivos
10	4,4002	5,4083e-06	Hipótese nula foi negada
11	0,10671	0.45751	Resultados inconclusivos
12	1,8823	0,029899	Hipótese nula foi negada
13	3,1931	0,00070389	Hipótese nula foi negada
14	22,7194	0	Hipótese nula foi negada
15	20,4042	0	Hipótese nula foi negada
16	6,4697	4,9091e-11	Hipótese nula foi negada
17	17,8993	0	Hipótese nula foi negada
18	6,3395	1,1528e-10	Hipótese nula foi negada

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

Diferentemente do visto nos níveis superiores, para as Subáreas os testes 1, 2 e 3 tiveram a hipótese nula negada, o que implica que tanto a representação numérica por TF-IDF, quanto a representação numérica usando 300 tópicos e o uso de informação de hierarquia são melhores do que o *baseline*.

Existem evidências que um número maior de tópicos na representação numérica por tópicos melhora o desempenho dos algoritmos (teste 4), assim como traduzir o texto dos títulos da produção científica dos pesquisadores melhora a acurácia dessa representação (testes 5 e 6). Além disso, a representação numérica baseada em 300 tópicos é melhor do que a representação por TF-IDF (teste 7).

Os testes 8 e 9 mostram que não têm como se tirar uma conclusão sobre qual das métricas de rede social é melhor. A abordagem numérica com 300 tópicos tem um desempenho melhor do que a por TF-IDF quando ambas adotam a métrica de rede social V1 (teste 10). Os resultados inconclusivos do teste 11 fazem com que não se seja possível afirmar se a métrica de rede social V1 melhora o resultado da representação numérica por

Quadro 7 – Quadro de hipóteses alternativas corretas - Subáreas

Teste	**Hipóteses Alternativas**
1	Representação numérica por TF-IDF é melhor do que o Baseline
2	Representação numérica por 300 tópicos é melhor do que o Baseline
3	Usar hierarquia de classes é melhor do que o Baseline
4	Usar 300 tópicos é melhor do que usar 150 tópicos na representação por tópicos
5	Texto traduzido é melhor do que ele sem traduzir na representação por 150 tópicos
6	Texto traduzido é melhor do que ele sem traduzir na representação por 300 tópicos
7	Representação numérica por 300 tópicos é melhor do que a por TF-IDF
10	Representação numérica por 300 tópicos com a métrica de rede social V1 é melhor do que a representação numérica por TF-IDF com a métrica de rede social V1
12	Representação numérica por 300 tópicos com a métrica de rede social V1 é melhor do que só a representação numérica por 300 tópicos
13	Usar informação de hierarquia (TF-IDF) é melhor do que só usar o TF-IDF
14	Usar informação de hierarquia (150 tópicos) é melhor do que só usar 150 tópicos
15	Usar informação de hierarquia (300 tópicos) é melhor do que só usar 300 tópicos
16	Usar informação de hierarquia (300 tópicos) é melhor do que usar informação de hierarquia (TF-IDF)
17	Usar informação de hierarquia (TF-IDF) é melhor do que usar a métrica de rede social V1
18	Usar informação de hierarquia (300 tópicos) é melhor do que usar informação de hierarquia (150 tópicos)

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

TF-IDF ou não. Porém, os resultados do teste 12 mostram que a representação numérica por tópicos tem um desempenho melhor quando a métrica de rede social V1 é utilizada.

Os testes 13, 14 e 15 mostram que usar a informação de hierarquia das classes melhora o desempenho dos algoritmos, sendo que a melhor representação dentre elas é a que adota a representação numérica com 300 tópicos (teste 16). Além disso, o uso da informação de hierarquia das classes possui resultados melhores do que usar a métrica de rede social V1 (teste 17).

7 Conclusão

Este projeto visou a estudar e avaliar diversas técnicas de inferência das áreas de atuação de pesquisadores usando a Plataforma Lattes como *gold standard* para fazer a avaliação do seu desempenho.

O objetivo do trabalho foi combinar diferentes características de forma a se avaliar a influência destas na acurácia da solução e estabelecer assim uma nova solução mais acurada do que as correlatas, ao menos ao se considerar a inferência da área de atuação dos pesquisadores cujos currículos estão cadastrados na Plataforma Lattes.

Assim, alguns dos objetivos específicos foram: avaliar se o enriquecimento dos dados melhora o desempenho dos algoritmos de classificação testados, além de analisar a contribuição de fatores como métricas de redes sociais, idioma dos títulos e a própria estrutura hierárquica das áreas de atuação no desempenho dos algoritmos. Todos esses objetivos foram cumpridos e algumas considerações podem ser feitas sobre os resultados.

A avaliação da inferência da área de atuação de um pesquisador pôde ser feita ao utilizar os diversos classificadores produzidos para os testes. Observou-se que, no geral, os resultados obtidos neste trabalho foram superiores aos presentes na literatura, em especial em relação ao *baseline* adotado (MIYATA; KANO; DIGIAMPIETRI, 2013), sobretudo nas representações numéricas por TF-IDF e por tópicos, cujos resultados foram consideravelmente melhores.

Nos testes desse trabalho foi constatado que o enriquecimento dos títulos da produção científica dos pesquisadores (artigos, orientações e projetos) não melhora o desempenho dos algoritmos, uma vez que o problema de inferência das áreas de atuação de pesquisadores tem classes que possuem sobreposição ou fronteiras não muito bem definidas umas com as outras. Portanto, o enriquecimento utilizado acaba adicionando ambiguidade em tais classes, diminuindo a acurácia total dos algoritmos.

A contribuição das métricas de redes sociais das áreas de atuação dos vizinhos (V1) e também da vizinhança nível dois (V2) foi avaliada, na qual ao se considerar somente as acurácias, os resultados obtidos foram superiores as suas contrapartes sem adotar as métricas de redes sociais. Ou seja, os conjuntos TF-IDF+VX/LanguageTfidfVX tiveram resultados superiores ao TF-IDF como representação numérica, bem como o Wiki-VX-Y teve acurácias melhores que o Wiki-Y.

Contudo, testes estatísticos aplicados sobre os resultados apresentam que, em alguns casos (por exemplo, no conjunto LanguageTfidfV1 comparado com o TF-IDF como representação numérica) não é possível se tirar alguma conclusão sobre qual abordagem é melhor: se adotar métricas de redes sociais melhora de fato ou não o desempenho. Esse tipo de situação torna inviável afirmar com veemência que as métricas de redes sociais melhoram o desempenho dos algoritmos em abordagens diversas, porém como existem casos que de fato há uma melhoria, pode-se afirmar que as métricas de redes sociais podem melhorar os classificadores (o que não implica que irão melhorar de fato). Além disso, os testes estatísticos provam que não é possível se afirmar qual das duas métricas (V1 ou V2) é melhor, uma vez que todos eles tiveram resultados inconclusivos.

O idioma dos títulos tem influência positiva sobre os algoritmos somente em casos específicos, pois só trás melhorias significativas (de acordo com os testes estatísticos) nas representações numéricas por tópicos, piorando o resultado dos algoritmos nas demais abordagens. Isso pode ser parcialmente explicado devido ao conjunto de dados externo possuir texto majoritariamente em português, tornando assim um texto traduzido mais adequado para fazer o mapeamento. Adicionalmente, o uso de técnicas mais complexas como as que adotam as métricas de rede social pode já ser capaz de capturar toda a contribuição que a influência dos idiomas traria para o problema (tornando essa informação desnecessária), porém cabendo a um trabalho futuro fazer um estudo sobre essa hipótese.

Duas formas de hierarquia também foram avaliadas: uma fazendo uma classificação em dois níveis no primeiro nível da taxonomia (Grandes Áreas) e outra usando as informações do nível superior para melhorar o desempenho dos classificadores para os dois níveis restantes (Áreas e Subáreas). No geral, apenas o uso de informação de hierarquia obteve o melhor desempenho nos seus respectivos níveis, cujos testes estatísticos confirmaram que ela é a abordagem que têm a melhor acurácia dentre todas as abordagens testadas por este trabalho. Destaca-se que para as Áreas e as Subáreas assumiu-se que a informação do nível superior estava disponível, o que pode não ser realidade em alguns problemas de inferência. A classificação em dois níveis obteve resultados análogos a não adotar a mesma, fazendo com que adotar essa abordagem seja desnecessário pois não há melhorias de desempenho.

Possíveis trabalhos futuros envolvem aplicar e testar as abordagens propostas em um conjunto de dados que não se limite somente aos bolsistas produtividade, podendo inclusive utilizar dados internacionais. Adicionalmente, os resultados da inferência poderiam ser

utilizados para sugerir atualizações no cadastro de um dado pesquisador. Além disso, um potencial trabalho futuro envolve propor uma abordagem que combine informações de diferentes métricas de redes sociais com a informação de hierarquia das classes, de forma a melhorar ainda mais a acurácia obtida neste trabalho. Outro trabalho envolve avaliar a escalabilidade nas abordagens propostas, de forma a estudar se o desempenho delas continua o mesmo conforme a complexidade do problema aumenta.

Referências[1]

AGGARWAL, C. C.; ZHAI, C. An introduction to text mining. In: _____. *Mining Text Data*. [S.l.]: Springer US, 2012. p. 1–10. Citado 2 vezes nas páginas 17 e 18.

BLEI, D. M.; MCAULIFFE, J. D. Supervised topic models. In: *In preparation*. [S.l.]: MIT Press, 2008. Citado na página 19.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 993–1022, 2003. Citado 3 vezes nas páginas 19, 20 e 21.

CANIBANO, C.; BOZEMAN, B. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. *Research Evaluation*, v. 18, n. 2, p. 86–94, 2009. Citado na página 14.

CHAGAS, F. M.; PEREZ-ALCAZAR, J. J.; DIGIAMPIETRI, L. A. *Algoritmo de classificação de especialistas em áreas na base de currículos Lattes*. [S.l.: s.n.], 2015. v. 21. 119-139 p. (Em Questão, v. 21). Citado na página 15.

CHEN, Y. H.; LI, S. F. Using latent dirichlet allocation to improve text classification performance of support vector machine. In: *2016 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.: s.n.], 2016. p. 1280–1286. Citado 5 vezes nas páginas 7, 24, 31, 32 e 33.

DIGIAMPIETRI, L.; MARUYAMA, W. Predição de novas coautorias na rede social acadêmica dos programas brasileiros de pós-graduação em ciência da computação. In: *CSBC 2014 - BraSNAM*. [S.l.: s.n.], 2014. Citado na página 15.

FONSECA, F.; DIGIAMPIETRI, L. A. Análise da relação entre obtenção de bolsas de produtividade do cnpq e medidas bibliométricas e de análise de redes sociais. In: *V Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2016)*. [S.l.: s.n.], 2016. Citado 4 vezes nas páginas 15, 39, 40 e 41.

GABRILOVICH, E.; MARKOVITCH, S. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*. [S.l.]: AAAI Press, 2006. (AAAI'06), p. 1301–1306. Citado na página 36.

GABRILOVICH, E.; MARKOVITCH, S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. [S.l.]: Morgan Kaufmann Publishers Inc., 2007. (IJCAI'07), p. 1606–1611. Citado 4 vezes nas páginas 24, 35, 36 e 45.

GRIFFITHS, T.; STEYVERS, M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, v. 101, n. SUPPL. 1, p. 5228–5235, 2004. Citado na página 21.

JR., C. S.; FREITAS, A. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, v. 22, n. 1-2, p. 31–72, 2011. Citado 2 vezes nas páginas 47 e 48.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

KATSURAI, M.; OHMUKAI, I.; TAKEDA, H. Topic representation of researchers' interests in a large-scale academic database and its application to author disambiguation. *IEICE Transactions on Information and Systems*, E99D, n. 4, p. 1010–1018, 2016. Citado 3 vezes nas páginas 23, 25 e 26.

LI, X. et al. A service mode of expert finding in social network. In: *2013 International Conference on Service Sciences (ICSS)*. [S.l.: s.n.], 2013. p. 220–223. Citado na página 36.

MIYATA, B. K. O.; KANO, V. Y.; DIGIAMPIETRI, L. A. Combinando mineração de textos e análise de redes sociais para a identificação das áreas de atuação de pesquisadores. In: *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013)*. [S.l.: s.n.], 2013. Citado 15 vezes nas páginas 24, 36, 39, 43, 44, 46, 51, 53, 54, 55, 57, 65, 66, 75 e 88.

NAVEED, N.; SIZOV, S.; STAAB, S. Att: Analyzing temporal dynamics of topics and authors in social media. In: *Proceedings of the 3rd International Web Science Conference, WebSci 2011*. [S.l.: s.n.], 2011. Citado 5 vezes nas páginas 19, 22, 23, 24 e 25.

PAUL, M.; GIRJU, R. Topic modeling of research fields: An interdisciplinary perspective. In: *International Conference Recent Advances in Natural Language Processing, RANLP*. [S.l.: s.n.], 2009. p. 337–342. Citado 2 vezes nas páginas 28 e 29.

PHAN, X.-H.; NGUYE, C.-T. Gibbslda++: A c/c++ implementation of latent dirichlet allocation (lda). 2007. Citado 2 vezes nas páginas 33 e 43.

PHAN, X.-H.; NGUYEN, L.-M.; HORIGUCHI, S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: *Proceedings of the 17th International Conference on World Wide Web*. [S.l.]: ACM, 2008. (WWW '08), p. 91–100. Citado 6 vezes nas páginas 7, 24, 33, 34, 39 e 42.

TANG, J. et al. A combination approach to web user profiling. *ACM Trans. Knowl. Discov. Data*, ACM, v. 5, n. 1, p. 2:1–2:44, 2010. Citado na página 14.

VO, D.; OCK, C. Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications*, v. 42, n. 3, p. 1684–1698, 2015. Citado 12 vezes nas páginas 20, 24, 29, 30, 31, 33, 34, 39, 44, 45, 51 e 58.

WANG, Z.; MA, L.; ZHANG, Y. A hybrid document feature extraction method using latent dirichlet allocation and word2vec. In: *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*. [S.l.: s.n.], 2016. p. 98–103. Citado 3 vezes nas páginas 24, 34 e 35.

WASSERSTEIN, R. L.; LAZAR, N. A. The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, Taylor & Francis, v. 70, n. 2, p. 129–133, 2016. Citado na página 73.

XU, S. et al. *Author-topic over time (AToT): A dynamic users' interest model*. [S.l.: s.n.], 2014. v. 274 LNEE. 239-245 p. (Lecture Notes in Electrical Engineering, v. 274 LNEE). Citado 3 vezes nas páginas 23, 24 e 25.

XU, Z. et al. Discovering user interest on twitter with a modified author-topic model. In: *Proceedings - 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011*. [S.l.: s.n.], 2011. v. 1, p. 422–429. Citado 3 vezes nas páginas 23, 26 e 27.

Apêndice A – Informações adicionais sobre os conjuntos de dados

Neste apêndice, informações adicionais sobre os conjuntos de dados utilizados nessa dissertação são apresentadas. Essas informações são compostas pela distribuição de classes de cada um dos problemas tratados (Grandes Áreas, Áreas e Subáreas), qual classe é a majoritária e a proporção da mesma com relação ao número total de instâncias em cada conjunto.

1 Grandes Áreas

O problema das Grandes Áreas possui 8 classes com 9.351 instâncias. A tabela 40 apresenta a distribuição de classes para as Grandes Áreas. A classe majoritária é a classe Ciências Exatas e da Terra, que representa $22,41\%$ de todo o conjunto de dados e tem 2.096 instâncias.

Tabela 40 – Distribuição de classes - Grandes Áreas

Classe	**Número de instâncias**
Ciências agrarias	1033
Ciências biológicas	1602
Ciências da saúde	1056
Ciências exatas e da terra	2096
Ciências humanas	1383
Ciências sociais aplicadas	615
Engenharias	1136
Linguística letras e artes	430

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

2 Áreas

O problema das Áreas possui 75 classes com 6.792 instâncias. A tabela 41 mostra a distribuição de classes para as Áreas. A classe majoritária é a classe Física, que representa $9,58$ % de todo o conjunto de dados e tem 651 instâncias.

Tabela 41 – Distribuição de classes - Áreas

Classe	**Número de instâncias**
Administração	109
Agronomia	377
Antropologia	102
Arqueologia	13
Arquitetura e urbanismo	51
Artes	62
Astronomia	79
Biofísica	18
Biologia geral	2
Bioquímica	84
Botânica	83
Ciência da computação	264
Ciência da informação	17
Ciência e tecnologia de alimentos	76
Ciência politica	82
Ciências ambientais	1
Comunicação	50
Demografia	6
Desenho industrial	7
Direito	51
Ecologia	73
Economia	165
Educação	203
Educação física	33
Enfermagem	113
Engenharia aeroespacial	7
Engenharia agrícola	29
Engenharia biomédica	13
Engenharia civil	121
Engenharia de materiais e metalúrgica	105
Engenharia de minas	5
Engenharia de produção	46
Engenharia de transportes	17
Engenharia elétrica	199
Engenharia mecânica	110
Engenharia naval e oceânica	5
Engenharia nuclear	14
Engenharia química	100
Engenharia sanitária	32
Farmácia	57
Farmacologia	63
Filosofia	101
Física	651
Fisiologia	44
Fisioterapia e terapia ocupacional	25
Fonoaudiologia	26

Continua na próxima página

Tabela 41 – continuação da distribuição de classes - Áreas

Classe	Número de instâncias
Genética	89
Geociências	7
Geografia	50
História	165
Imunologia	53
Letras	158
Linguística	115
Matemática	274
Medicina	336
Medicina veterinária	137
Microbiologia	43
Morfologia	30
Nutrição	16
Oceanografia	26
Odontologia	138
Parasitologia	23
Planejamento urbano e regional	9
Probabilidade e estatística	50
Psicologia	187
Química	479
Recursos florestais e engenharia florestal	65
Recursos pesqueiros e engenharia de pesca	7
Saúde coletiva	78
Serviço social	24
Sociologia	107
Teologia	4
Turismo	2
Zoologia	65
Zootecnia	164

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017

3 Subáreas

O problema das Subáreas possui 484 classes com 4.060 instâncias. A tabela 42 apresenta a distribuição de classes para as Subáreas. A classe majoritária é a classe Física da Matéria Condensada, que representa $6,37$ % de todo o conjunto de dados e tem 259 instâncias.

Tabela 42 – Distribuição de classes - Subáreas

Classe	Número de instâncias
Acústica	1
Administração	2
Administração de empresas	55
Administração educacional	1
Administração em enfermagem	2
Administração financeira	1
Administração publica	1
Agrometeorologia	8
Álgebra	28
Analise	49
Analise clinicas	1
Analise do discurso	2
Analise e controle de medicamentos	5
Analise nutricional de população	8
Analise toxicológica	7
Analises clinicas	2
Anatomia	3
Anatomia patológica e patologia clínica	17
Antropologia da religião	2
Antropologia das populações afro-brasileiras	2
Antropologia rural	4
Antropologia social	1
Antropologia urbana	24
Antropologia visual	1
Aplicações de radioisótopos	4
Aprendizado de máquina	2
Aquicultura	6
Aquisição da linguagem	1
Áreas clássicas de fenomenologia e suas aplicações	2
Arqueologia histórica	2
Arqueologia pré-histórica	6
Arquivologia	3
Artes do vídeo	1
Artes plásticas	4
Artes visuais	1
Astrofísica	1
Astrofísica do sistema solar	1
Astrofísica estelar	12
Astrofísica extragaláctica	18
Astronomia de posição e mecânica celeste	4
Atividade física e saúde	4
Audiologia	8
Bacteriologia	1
Biblioteconomia	5
Bioengenharia	6
Bioética	2

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Biofísica celular	3
Biofísica das radiações	1
Biofísica molecular	6
Biologia celular	4
Biologia da conservação	2
Biologia do desenvolvimento	1
Biologia e fisiologia dos microrganismos	16
Biologia molecular	16
Biomateriais	1
Biomecânica	5
Bioquímica clínica	1
Bioquímica da nutrição	3
Bioquímica de plantas	1
Bioquímica dos microrganismos	3
Biotecnologia	4
Botânica aplicada	2
Bromatologia	1
Cardiologia	1
Cariologia	1
Catalisadores	1
Catalise	1
Ciência de alimentos	25
Ciência do solo	94
Ciências ambientais	1
Ciências contábeis	14
Ciências do esporte	1
Cinema	4
Circuitos elétricos magnéticos e eletrônicos	5
Cirurgia	35
Cirurgia bucomaxilofacial	3
Citologia e biologia celular	10
Clínica e cirurgia animal	29
Clinica medica	136
Clínica odontológica	8
Clínica veterinária	1
Componentes da dinâmica demográfica	1
Comportamento animal	5
Comportamento motor	2
Comportamento organizacional	1
Comportamento político	6
Comunicação visual	2
Conservação	1
Conservação da biodiversidade	2
Conservação da natureza	3
Construção civil	19
Construções rurais e ambiência	2
Corrosão	1

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Cosmologia	1
Crescimento e desenvolvimento econômico	1
Crescimento flutuações e planejamento econômico	1
Cultura brasileira	1
Currículo	7
Dança	3
Desnutrição e desenvolvimento fisiológico	2
Dietética	2
Dinâmica de voo	2
Direito agrário direito ambiental direitos humanos	1
Direito ambiental	1
Direito público	4
Direitos especiais	1
Disfunção temporomandibular	1
Distúrbios da comunicação humana	1
Distúrbios da linguagem	1
Divulgação cientifica	2
Ecologia aplicada	5
Ecologia de ecossistemas	22
Ecologia dos animais domésticos e etologia	1
Ecologia microbiana	1
Ecologia teórica	8
Economia do bem-estar social	3
Economia do trabalho	4
Economia dos recursos humanos	4
Economia industrial	1
Economia internacional	4
Economia monetária e fiscal	4
Economia política	1
Economia regional e urbana	3
Economias agraria e dos recursos naturais	3
Educação a distância	2
Educação ambiental	2
Educação de jovens e adultos	1
Educação e trabalho	2
Educação especial	2
Educação física	3
Educação matemática	1
Educação superior	3
Eletrônica de potência	2
Eletrônica industrial sistemas e controles eletrônicos	33
Embriologia	2
Endodontia	5
Energias renováveis	1
Energização rural	1
Enfermagem de doenças contagiosas	1
Enfermagem de saúde publica	20

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Enfermagem fundamental	1
Enfermagem medicocirurgica	30
Enfermagem obstétrica	8
Enfermagem pediátrica	5
Enfermagem psiquiátrica	6
Engenharia acústica	1
Engenharia ambiental	2
Engenharia de agua e solo	15
Engenharia de alimentos	10
Engenharia de processamento de produtos agrícolas	3
Engenharia de reatores	1
Engenharia de software	2
Engenharia econômica	2
Engenharia hidráulica	2
Engenharia medica	1
Engenharia térmica	11
Ensino aprendizagem	12
Ensino de enfermagem	2
Ensino superior	3
Entomologia	1
Entomologia e malacologia de parasitos e vetores	7
Enzimologia	5
Epidemiologia	24
Epistemologia	6
Estado e governo	4
Estatística	11
Estatística aplicada a engenharia	1
Estruturas	31
Estruturas aeroespaciais	1
Estudos organizacionais	1
Ética	2
Etnofarmacologia	1
Etnologia indígena	7
Farmacognosia	9
Farmacologia	2
Farmacologia autonômica	1
Farmacologia bioquímica e molecular	4
Farmacologia da inflamação e da dor	1
Farmacologia geral	8
Farmacotecnia	10
Fenômenos de transporte	19
Filogenia	1
Filosofia contemporânea	1
Filosofia da educação	1
Filosofia política	1
Física atômica e molecular	21
Física biomolecular	1

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Física da matéria condensada	259
Física das partículas elementares e campos	79
Física dos fluidos física de plasmas e descargas elétricas	7
Física geral	22
Física medica	2
Física nuclear	17
Físico-química	61
Fisiologia comparada	1
Fisiologia da reprodução	1
Fisiologia de órgãos e sistemas	40
Fisiologia de sementes	1
Fisiologia do esforço	14
Fisiologia do exercício	8
Fisiologia geral	1
Fisiologia respiratória	1
Fisiologia vegetal	21
Fisioterapia	4
Fisioterapia pulmonar	1
Fisioterapia respiratória	1
Fitossanidade	83
Fitotecnia	87
Floricultura parques e jardins	1
Fontes renováveis de energia	1
Formação de professores	1
Fruticultura	1
Fundamentos da educação	27
Fundamentos da sociologia	4
Fundamentos de arquitetura e urbanismo	6
Fundamentos do planejamento urbano e regional	1
Fundamentos do serviço social	7
Fundamentos e critica das artes	5
Fundamentos e medidas da psicologia	6
Gênero e educação	1
Genética	1
Genética animal	15
Genética de populações	2
Genética e melhoramento de plantas	1
Genética e melhoramento dos animais domésticos	8
Genética humana e medica	26
Genética molecular e de microorganismos	8
Genética quantitativa	1
Genética vegetal	8
Geografia física	1
Geografia humana	26
Geografia regional	1
Geologia	4
Geometria e topologia	68

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Geotécnica	29
Gerencia de produção	8
Gerenciamento em enfermagem	1
Gestão ambiental	1
Gestão da informação	1
Ginecologia	1
Helmintologia de parasitos	4
Hematologia	1
Herpetologia	1
Hidrodinâmica de navios e sistemas oceânicos	2
Histologia	6
História ambiental	1
História antiga e medieval	11
História cultural	1
História da África	2
História da américa	5
História da arte	1
História da filosofia	18
História da literatura	1
História das ciências	5
História das teologias e religiões	2
História do Brasil	40
História intelectual	1
História moderna	2
História moderna e contemporânea	3
História oral	1
Implantodontia	1
Imunofarmacologia	4
Imunogenetica	1
Imunologia	2
Imunologia aplicada	3
Imunologia celular	25
Imunologia molecular	1
Imunoquimica	1
Infectologia	1
Informação quântica	3
Infraestrutura de transportes	2
Inspeção de produtos de origem animal	2
Instalações e equipamentos metalúrgicos	1
Inteligência artificial	3
Inteligência computacional	2
Inteligência organizacional	1
Interdisciplinar	1
Irrigação e drenagem	2
Jornalismo e editoração	3
Lavra	1
Limnologia	1

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Língua portuguesa	2
Linguagem	4
Línguas estrangeiras modernas	1
Linguística aplicada	21
Linguística histórica	2
Literatura	1
Literatura brasileira	6
Literatura comparada	5
Literatura infanto-juvenil	1
Literaturas clássicas	3
Literaturas estrangeiras modernas	2
Logica	3
Manejo florestal	11
Maquinas e implementos agrícolas	5
Marketing	1
Matemática	1
Matemática aplicada	43
Matemática da computação	5
Materiais e processos para Engenh. Aeronáutica e aeroespacial	3
Materiais naometalicos	31
Materiais odontológicos	6
Mecânica dos corpos sólidos elásticos e plásticos	2
Mecânica dos fluidos	1
Mecânica dos sólidos	17
Medicina veterinária preventiva	31
Medidas elétricas magnéticas e eletrônicas instrumentação	3
Melhoramento de plantas	1
Memoria social	2
Metabolismo e bioenergética	13
Metafisica	1
Metalurgia de transformação	2
Metalurgia extrativa	4
Metalurgia física	16
Metodologia cientifica	2
Metodologia da pesquisa	1
Metodologia de pesquisa	2
Metodologia e técnicas da computação	73
Métodos quantitativos em economia	11
Micologia	1
Microbiologia aplicada	14
Morfologia	1
Morfologia dos grupos recentes	7
Morfologia vegetal	12
Movimentos sociais	1
Multidisciplinar	1
Museologia	1
Musica	24

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Mutagênese	2
Nanocompositos	1
Nanotecnologia	2
Neolinguística	2
Neurociências	5
Neurofarmacologia	1
Neurofisiologia	1
Neurologia	2
Neuropsicofarmacologia	13
Neuroquímica	1
Nutrição clínica	1
Nutrição e alimentação animal	24
Oceanografia biológica	8
Oceanografia física	10
Oceanografia geológica	2
Oceanografia química	1
Odontologia social e preventiva	2
Odontopediatria	8
Oftalmologia	2
Oncologia	1
Operações de transportes	2
Operações industriais e equipamentos para engenh. Química	7
Optica quântica	2
Organizações	1
Ortodontia	6
Otimização	1
Otimização combinatória	1
Outras literaturas vernáculas	4
Outras sociologias especificas	13
Paisagismo	3
Paleobotânica	1
Parasitologia veterinária	1
Pastagem e forragicultura	14
Patologia	3
Patologia animal	12
Patologia bucal	8
Patologia clínica	3
Patologia geral	1
Pediatria	3
Periodontia	12
Pesquisa operacional	20
Planejamento de fármacos	1
Planejamento de transportes	7
Planejamento e avaliação educacional	6
Planejamento urbano e regional	2
Política de saúde	1
Política educacional	1

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Política internacional	11
Política pública e população	1
Politicas públicas	4
Políticas sociais	1
Probabilidade	9
Probabilidade e estatística aplicadas	19
Processamento de sinais	2
Processos de fabricação	12
Processos estocásticos	2
Processos industriais de engenharia química	12
Produção animal	8
Produção vegetal	1
Produtos naturais	2
Programação visual	2
Projeto de arquitetura e urbanismo	6
Projetos de maquinas	1
Protozoologia de parasitos	11
Psicanalise	2
Psicolinguística	4
Psicologia clínica	1
Psicologia cognitiva	2
Psicologia comunitária	1
Psicologia da saúde	1
Psicologia do desenvolvimento humano	10
Psicologia do ensino e da aprendizagem	3
Psicologia do trabalho e organizacional	5
Psicologia educacional	1
Psicologia experimental	7
Psicologia fisiológica	3
Psicologia social	25
Psiquiatria	29
Química	2
Química analítica	102
Química computacional	1
Química de macromoléculas	23
Química farmacêutica	1
Química inorgânica	49
Química medicinal	2
Química orgânica	106
Química supramolecular	1
Química verde	1
Radiologia medica	9
Radiologia odontológica	2
Recursos hídricos	5
Recursos pesqueiros marinhos	1
Redes de computadores	5
Redes de computadores e sistemas distribuídos	2

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Relações internacionais	2
Relações públicas e propaganda	1
Reprodução animal	32
Reprodução de peixes	1
Resíduos sólidos	1
Ressonância magnética nuclear	1
Saneamento ambiental	4
Saneamento básico	3
Saúde coletiva	1
Saúde da mulher	1
Saúde materno infantil	30
Saúde pública	20
Semiótica	1
Serviço social aplicado	10
Serviços urbanos e regionais	1
Silvicultura	12
Sinalização celular	1
Sistemas de computação	41
Sistemas dinâmicos	4
Sistemas elétricos de potência	38
Sistemas inteligentes	1
Sociolinguística e dialetologia	2
Sociologia da cultura	2
Sociologia da educação	1
Sociologia da saúde	2
Sociologia do conhecimento	1
Sociologia econômica	1
Sociologia política	3
Sociologia rural	4
Sociologia urbana	2
Taxonomia dos grupos recentes	18
Taxonomia vegetal	14
Teatro	10
Técnicas e operações florestais	2
Tecnologia de alimentos	10
Tecnologia de arquitetura e urbanismo	4
Tecnologia dos reatores	7
Tecnologia e utilização de produtos florestais	11
Tecnologia química	5
Telecomunicações	40
Teologia sistemática	1
Teoria da computação	17
Teoria da comunicação	17
Teoria da informação	1
Teoria das organizações	1
Teoria do direito	11
Teoria e analise linguística	30

Continua na próxima página

Tabela 42 – continuação da distribuição de classes - Subáreas

Classe	Número de instâncias
Teoria e filosofia da história	1
Teoria e método em arqueologia	1
Teoria econômica	10
Teoria literária	7
Teoria política	2
Teoria social	1
Terminologia	1
Tópicos específicos de educação	13
Toxicologia	7
Tratamento de águas de abastecimento e residuarias	3
Tratamento de minérios	1
Tratamento e prevenção psicológica	13
Virologia	3
Voz	1
Zoologia aplicada	3

Fonte: Felipe Penhorate Carvalho da Fonseca, 2017