IMPACTO DE COMUNIDADES SOCIAIS ONLINE NO APRENDIZADO DE IDIOMAS

RAFAEL SALES MEDINA FERREIRA

IMPACTO DE COMUNIDADES SOCIAIS ONLINE NO APRENDIZADO DE IDIOMAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Fabricio Murai Ferreira
Coorientador: Ana Paula Couto da Silva

Belo Horizonte
Maio de 2019

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Ferreira, Rafael Sales Medina.

F383i Impacto de comunidades sociais online no aprendizado de idiomas / Rafael Sales Medina Ferreira — Belo Horizonte, 2019.
xxii, 52. il.; 29 cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Fabrício Murai Ferreira
Coorientadora: Ana Paula Couto da Silva

1. Computação – Teses. 2. Aprendizado de máquinas. 3. Redes sociais. 4. Redes complexas. 5. Aprendizado de idiomas assistido por computador. I. Orientador. I. Orientadora. III. Título.

CDU 519.6*82.10 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Impacto de Comunidades Sociais Online no Aprendizado de Idiomas

RAFAEL SALES MEDINA FERREIRA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. FABRÍCIO MURAI FERREIRA - Orientador
Departamento de Ciência da Computação - UFMG

PROFA. ANA PAULA COUTO DA SILVA - Coorientadora
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA
Departamento de Ciência da Computação - UFMG

PROF. DANIEL SADOC MENASCHE
Departamento de Ciência da Computação - UFRJ

Belo Horizonte, 2 de Maio de 2019.

Dedico este trabalho a todos que acreditaram em mim, mesmo quando eu mesmo não acreditava.

Agradecimentos

Agradeço aos meus pais, Denise e Ernani, por serem a base de tudo que sou hoje. Muito obrigado pelo apoio, carinho e amor incondicional. Sem vocês não estaria onde estou. Extensivo a Douglas e Alexandre, meus irmãos e companheiros desde sempre.

Agradeço aos meus incríveis orientadores, Ana e Fabrício, por me conduzirem e acompanharem por toda essa jornada. Muito obrigado pelo apoio, atenção, confiança, disponibilidade e por todos os conselhos ao longo do mestrado.

À minha tia Latife, a quem serei eternamente grato por me acolher por tanto tempo e por toda sua atenção e carinho. Aos meus avós, Bilá, Celina, Jayme e Miguel, por tanto amor e por acreditarem sempre no meu potencial.

A todos os meus amigos e familiares, muito obrigado por estarem sempre comigo! Mas especialmente, muito obrigado pela paciência!

Ao Lucas, pelo companheirismo, apoio e paciência, principalmente ao longo desses dois últimos anos. E à Andrea, por sua atenção e cuidado.

Thiago e Maria, por compartilharem comigo alguns momentos e opiniões ao longo dos últimos anos, e por terem me ensinado tanto. À Marina, por estar do meu lado desde sempre, e para sempre.

Glivia, por ser uma das minhas grandes incentivadoras e por ter me ajudado a chegar até aqui.

Bárbara e Gustavo, grandes amigos que o DCC me trouxe, muito obrigado por compartilharem comigo momentos bons, e alguns momentos tensos também.

Mateus e Diogo, por estarem sempre ao meu lado para me ouvir, aconselhar e ajudar sempre que preciso.

Amanda, Lorena, Luiza, Caik, Rodrigo, Pedro, Gabriel e Maria, grandes amigos que estão comigo nas melhores e piores horas também, sempre trazendo palavras que tanto me ajudam.

Aos meus amigos do DataViva e dti, por me acolherem no meio dessa jornada e por tantos bons momentos ao longo dos meus dias. Agradeço também por confiarem no meu trabalho e me ensinarem tanto, principalmente ao longo do último ano.

Ao Tiago, por me ajudar a me encontrar e me entender, e a ser uma pessoa melhor comigo mesmo a cada dia.

Agradeço também ao CNPq, ao Departamento de Ciência da Computação e à Universidade Federal de Minas Gerais, extensivo a todos os excelentes profissionais que cruzaram meu caminho durante essa jornada.

"Every time that we run
We don't know what it's from
Now we finally slow down
We feel close to it
There's a change gonna come
I don't know where or when
But whenever it does
We'll be here for it"
(Lana Del Rey)

Resumo

O Reddit é uma rede social online em que usuários interessados em um mesmo tópico se inscrevem em comunidades (subreddits) onde compartilham conteúdo (e.g., links, texto, imagens) na forma de posts. Posts são, por sua vez, comentados e avaliados por outros usuários. Subreddits para aprendizado de idioma vem atraindo mais usuários de diferentes níveis de proficiência a cada ano, que interagem compartilhando dúvidas e dicas de como melhorar o aprendizado. É importante analisar o conteúdo compartilhado e compreender como se dão as interações entre usuários nessas redes para que seja possível projetar e implementar novas soluções para melhorar a experiência do usuário e consequentemente auxiliar no aprendizado. Neste trabalho, analisamos as redes para aprendizado de alemão, espanhol, francês e inglês, concluindo que usuários demonstram maior interesse na discussão do assunto do que nas interações interpessoais, como comprovado pela análise da rede de interação de cada uma das comunidades. Essa análise demonstrou que usuários não possuem laços muito fortes entre si, sendo que não há muitos pares conectados, e quando há interação entre dois usuários, ela se repete poucas vezes. Isso é corroborado por análises das métricas de centralidade de redes complexas, que demonstram que as redes de subreddits não apresentam estrutura semelhante à de redes sociais tradicionais. Além disso, analisamos as threads, onde percebe-se que as discussões em posts não se prolongam por muito tempo. Observa-se que geralmente a primeira publicação recebe muitas respostas, mas a discussão não se estende muito a partir dessas respostas. Muitas vezes as publicações são perguntas, e usuários respondem rapidamente a essas perguntas e a discussão termina. Utilizando o subreddit German, onde os usuários são aconselhados a informar seu nível de proficiência quando fazem uma publicação (post/comentário), utilizamos a ferramenta LIWC para extrair características dos textos deste idioma, o que permitiu observar que publicações de usuários de níveis diferentes de proficiência apresentam características textuais bem distintas. Esta observação nos levou a investigar o problema de se classificar a proficiência dos usuários a partir de suas publicações. Por meio de uma série de experimentos, demonstramos que classificadores que tratam as publicações no

Reddit como observações independentes, como encontrados na literatura, têm baixo desempenho. À vista disso, propomos um novo modelo, SEMPLICe (SEquential Model for Proficiency cLassifICation), que considera as características textuais e também o histórico de um usuário no subreddit para classificar sua proficiência ao longo do tempo. Baseado na suposição de que a proficiência é não decrescente desde que um usuário permaneça ativo, SEMPLICe alcança um F1 ponderado até 29,6% maior que os métodos clássicos. SEMPLICe utiliza programação dinâmica para obter complexidade linear no tamanho do histórico de cada usuário.

Palavras-chave: Aprendizado de Idiomas Assistido por Computador, Redes Sociais, Redes Complexas, Aprendizado de Máquina.

Abstract

Reddit is a social network where users interested in a common subject may subscribe to communities, known as subreddits, were they can share content such as links, text and images as posts. Other users in the community can, in turn, comment and rate such posts. Subreddits for second language learning have been drawing attention from groups of users of the most diverse proficiency levels, who use these communities for sharing questions and tips on improving their level. It is important to analyse the content shared on such communities and understand how users interact in order to guide the design of new tools that can improve language learning as well as user experience. In this project, we analyse subreddits for language learning, namely German, English, French and Spanish. We analyse the network of interactions in these subreddits and show that users are more focused on discussing the subject than on interpersonal interactions. This analysis also shows that most of the relationships between users are weak ties: when two users interact with each other, that interaction typically does not reoccur many times. This conclusion is corroborated by our analysis of centrality metrics of complex networks, which show that these networks do not share common features with traditional online social media. Moreover, we analyse threads and show that discussion topics do not have long reply threads. Instead, threads usually have many answers to the first post, none of those leading to longer discussions. Using subreddit German, where users are asked to inform their proficiency level, we use LIWC to extract linguistic features from posts published in this subreddit, which indicates that users with different proficiency levels write text with distinct textual features. This observation led us to investigate whether it is possible to categorize users proficiency based on their publications. Unfortunately, traditional classification models such as KNN and logistic regression result in low accuracy when predicting proficiency from textual features extracted from individual Reddit publications. To address this problem, we propose a new model, called SEMPLICe (SEquential Model for Proficiency cLassifICation), which considers both text features and users history of publications to classify proficiency level throughout time. Based on the premise that proficiency

levels do not decrease, as long as the users are active on the interaction, SEMPLICe improves the F1 metrics from classic methods up to 29.6%. SEMPLICe uses dynamic programming in order to obtain linear complexity on the users interaction line.

Keywords: Computer-Assisted Language Learning, Social Networks, Complex Networks, Machine Learning.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Ao decorrer das últimas duas décadas, uma área de estudos que vem recebendo crescente atenção por pesquisadores é a *Computer Assisted Language Learning* (CALL) [Levy, 1997], ou Aprendizado de Idiomas Apoiado por Computadores, em tradução livre. Essa área engloba quaisquer tipos de aplicações que possam auxiliar no aprendizado de idiomas estrangeiros.

Em particular, alguns estudos recentes na área têm analisado redes sociais online [Zourou, 2012], como a Livemocha, criada especificamente para discussões de alunos em relação a um idioma que estão aprendendo [Clark & Gruba, 2010]. Nesse contexto, uma rede social que ainda não foi estudada como apoio para aprendizado de idiomas é o Reddit.

O Reddit[1] é um site de fóruns com características de redes sociais. Seu nome tem como origem um jogo de palavras com a expressão inglesa "read it", que em tradução livre significa "lido" e a ideia é indicar que certo conteúdo foi "lido no Reddit". Esta rede permite aos usuários criar e compartilhar conteúdo em forma de *posts*, que podem ser respondidos por outros usuários por meio de comentários. Esses comentários, por sua vez, também podem ser respondidos com outros comentários, e essa troca de informação e conteúdo pode ser vista como uma árvore de discussão, conhecida como *thread*. Os usuários também podem decidir, por meio de votação, quais são os *posts* mais relevantes [Anderson, 2015].

O Reddit é dividido em subreddits, frequentemente chamados de comunidades, que são fóruns com temas específicos. Existem subreddits específicos para discutir programas de televisão[2], jogos de computador[3] e até mesmo para compartilhamento de

[1] http://www.reddit.com
[2] http://www.reddit.com/r/rupaulsdragrace
[3] http://www.reddit.com/r/thesims

tópicos relacionados à saúde, como dicas de emagrecimento[4], conforme estudado por Pappa et al. [2017]; Cunha et al. [2016], e suporte mútuo em relação a problemas de saúde mental, como estudados por De Choudhury & De [2014]; Silveira et al. [2018]; Fraga et al. [2018] ou vítimas de abuso sexual, como estudado por Andalibi et al. [2016].

Assim como em outras redes sociais, no Reddit também são encontradas comunidades voltadas para aprendizado de novas línguas. Existem comunidades específicas voltadas para usuários que estão aprendendo uma língua estrangeira, como alemão[5] e francês[6]. Esses subreddits permitem que pessoas com interesse em certo idioma interajam, compartilhando dúvidas, sugestões e dicas.

O Reddit se mostra como uma boa ferramenta de apoio para usuários com interesse no aprendizado de um idioma, dada a participação crescente nas comunidades e as possibilidades de interação que os subreddits proporcionam. Contudo, diferente de outras redes sociais cuja proposta é apenas o aprendizado de um ou mais idiomas, o Reddit é uma rede social de propósito mais geral, onde comunidades voltadas para este fim específico surgiram organicamente. Em vista disso, observa-se uma necessidade de análise mais profunda dessas redes, tendo como foco tanto as interações quanto as publicações em si. Essa análise tem como objetivo compreender melhor o funcionamento desses subreddits, o que consequentemente permite o aprimoramento e a criação de soluções para melhorar a experiência do usuário e auxiliar ainda mais no aprendizado de línguas estrangeiras.

Uma particularidade interessante de alguns desses subreddits é a utilização de *tags* (etiquetas) pelos usuários. O subreddit *French*, por exemplo, é voltado para o ensino de francês e define *tags* para cada tópico, de maneira a categorizar cada postagem por tema: discussões, mídia, recursos e conselhos. O subreddit *German*, voltado para discussões sobre o alemão, categoriza os usuários em: usuários que possuem alemão como língua nativa (colocando em suas publicações uma *tag* de cor verde); usuários que estão aprendendo (*tag* de cor azul); e usuários que são graduados em linguística (*tag* de cor dourada). Nesse mesmo subreddit, juntamente com a cor da *tag*, os usuários que estão aprendendo também recebem a recomendação de incluir uma *tag* textual, informando seu nível de proficiência no idioma, de maneira a auxiliar na interação com usuários de níveis diferentes.

A categorização de postagens e usuários dentro de cada subreddit pode influenciar a maneira como as pessoas interagem dentro do mesmo, podendo, por sua vez, impactar na forma como a rede de interações se organiza. Mesmo que a rede de alemão

[4]http://www.reddit.com/r/loseit
[5]http://www.reddit.com/r/German/
[6]http://www.reddit.com/r/French/

apresente muitos usuários com informação sobre proficiência no perfil, observamos que uma parcela também considerável dos usuários não apresenta essa informação. Essa constatação nos leva a questionar se é possível classificar a proficiência desses usuários de acordo com suas características textuais. A classificação dos textos é interessante porque permite aos usuários um acompanhamento mais próximo de sua evolução, especialmente para aqueles que inicialmente não sabem seu nível de proficiência.

Esta classificação automática poderia ser usada para aumentar o engajamento dos usuários e, por conseguinte, melhorar o aprendizado de idiomas. Por exemplo, poderia ser usada para informar ao usuário a melhoria no seu nível de proficiência, ou mesmo sugerir a um usuário participar de uma *thread* na qual estão interagindo usuários de nível maior.

Assim, definimos como objetivos principais para esta dissertação: (1) compreender e mensurar como redes sociais online são utilizadas como apoio por usuários para o aprendizado de idiomas estrangeiros e (2), definir modelos que permitam acompanhar a evolução da proficiência. Com base nesses objetivos, organizamos esta dissertação em torno de três questões de pesquisa, conforme apresentado na seção a seguir.

1.1 Questões de pesquisa

As questões de pesquisa que norteiam esta dissertação são:

1. **As interações entre usuários em um subreddit são semelhantes àquelas em uma rede social tradicional?**

   Desejamos identificar o objetivo dos usuários ao interagir em um subreddit: estão apenas procurando novos contatos para se conectar, estão simplesmente em busca de compartilhar e obter conhecimento, ou possuem ambos os objetivos?

2. **Como as publicações em um subreddit estão distribuídas em relação às threads?**

   Os subreddits são compostos por diversas *threads* onde usuários podem compartilhar conteúdo. Desejamos compreender como ocorre a interação dentro dessas threads, avaliando, por exemplo, se discussões longas são bastante frequentes e quais os assuntos mais atraem usuários.

3. **É possível classificar a evolução da proficiência em uma certa língua estrangeira a partir de publicações em uma rede social online?**

Sabendo que existe um grande número de usuários com níveis de proficiência distintos interagindo nas comunidades, investigamos se textos de usuários com níveis distintos apresentam características diferentes. Analisamos também a possibilidade de classificar a proficiência de um texto baseado nessas características, considerando ou não o histórico de publicações de um usuário.

1.2 Contribuições

Com a finalidade de responder às questões anteriores, realizamos diversas análises utilizando os dados dos principais subreddits voltados para aprendizado de idiomas. A partir delas surgiram as principais contribuições desta dissertação, a saber:

Modelagem e caracterização das interações. Para responder a primeira questão de pesquisa, caracterizamos as interações entre usuários dos subreddits a partir da modelagem da rede de usuários como um grafo. Nessa abordagem, cada usuário que publicou um *post* (uma publicação direta no site, iniciando uma nova *thread*) ou comentário (uma resposta a um *post* ou outro comentário) é representado por um nó. Uma aresta saindo de um usuário v para outro usuário u indica que v responde a uma atividade (um *post* ou comentário) de u. Essa ligação terá um peso w, igual ao número de vezes que v respondeu a um *post* ou comentário de u durante o período considerado. A partir dessa modelagem, foram calculadas as métricas relacionadas à rede, como *closeness* e coeficiente de clusterização.

Essas métricas foram, por sua vez, utilizadas para analisar os padrões de comportamento dos usuários que participam de um subreddit. Esta modelagem é utilizada em diversos trabalhos da literatura como realizada por Pappa et al. [2017]. Ela permitiu que se traçasse um perfil daqueles usuários mais ativos na rede, ou seja, dos usuários que mais fizeram *posts* ou comentaram publicações de outros.

Modelagem e caracterização das discussões em threads. Para responder à segunda questão de pesquisa, estudamos como as discussões entre usuários dentro de uma mesma *thread* ocorrem. Para isso, cada *thread* dos subreddits foi modelada como uma árvore, em que a publicação original é a raiz e os comentários representam as ramificações. Essa modelagem permitiu observar o comportamento dos usuários em relação às publicações, visto que analisamos a quantidade de respostas que um post recebe, por exemplo. Além disso, também observamos qual o tipo de publicação que recebe mais respostas dos usuários. As análises realizadas nesse trabalho permitem a compreensão da organização de comunidades do *Reddit* específicas para aprendizado de idiomas.

Extração de características textuais e avaliação de métodos clássicos de aprendizado supervisionado para classificação automática de proficiência nos subreddits. Nas duas primeiras análises, observamos a grande quantidade de usuários que informam seu nível de proficiência na rede voltada para o ensino de alemão, e como esses usuários interagem bastante entre si. Dessa maneira, como resposta à terceira questão de pesquisa, realizamos a análise textual utilizando a ferramenta LIWC (Linguistic Inquiry and Word Count), um dicionário que analisa textos e retorna as características dos mesmos (como quantidade e tamanho das palavras utilizadas por frase). Esses valores, associados com *tags* de proficiência que os próprios usuários incluem em suas publicações, permitiram verificar que existem diferenças entre os textos de usuários com diferentes níveis de proficiência.

Utilizamos as características textuais das publicações e seus rótulos para treinar os seguintes modelos de classificação: KNN, *Random Forest*, *Logistic Regression*, *Gradient Boosting* e XGBoost, utilizados com a finalidade de classificar outros textos quanto à sua proficiência. A aplicação desses modelos resultou em um valor de acurácia muito baixo, visto que esses modelos consideram cada publicação como um evento único.

Proposta de um modelo sequencial para classificação automática de proficiência. Como modelos que consideram cada publicação como evento isolado apresentam uma acurácia baixa para classificar a proficiência de um usuário, propomos um novo modelo que considera a sequência em que as publicações foram realizadas, baseado na proposição de que a proficiência em um idioma deve se manter constante ou crescente ao decorrer do tempo. Esse novo modelo, denominado SEMPLICe, considera a proficiência do usuário na publicação anterior àquela que está classificando para realizar a classificação da proficiência, além de utilizar a probabilidade da postagem atual apresentar cada nível de proficiência. Essas probabilidades são resultantes da aplicação dos métodos que apresentaram melhor acurácia na etapa anterior.

A classificação da proficiência das publicações do Reddit por meio do SEMPLICe possui alta acurácia, alcançando F_1 ponderado superior a 60% de acerto, valor que é comparável aqueles observados em trabalhos baseados em textos mais longos e mais formais, que oscilam em torno de 70% de acurácia [Crossley et al., 2012]. Isso demonstra que é possível categorizar publicações de usuários baseado em suas publicações. Dessa maneira, um usuário pode acompanhar a própria evolução da sua proficiência, sabendo o momento em que alcançou o próximo nível baseado em sua atividade e textos. Outro exemplo de aplicação do modelo é sua utilização por professores do idioma, que podem utilizar para acompanhar a evolução de alunos. O modelo pode também servir como base para a criação de aplicativos e mídias sociais específicos e mais eficazes para o aprendizado de línguas por meio da interação entre pessoas utilizando uma rede virtual.

1.3 Estrutura da dissertação

O restante desta dissertação está organizada da seguinte forma. O Capítulo 2 apresenta a revisão da literatura, envolvendo trabalhos sobre Computer Assisted Language Learning, Reddit e Classificação automática de proficiência, além de apresentar um comparativo deste trabalho com os trabalhos encontrados. O Capítulo 3 descreve detalhadamente a metodologia aplicada em cada etapa de trabalho. O Capítulo 4 apresenta o SEMPLICe, modelo proposto para classificar a proficiência de usuários da rede e principal contribuição desta dissertação. O Capítulo 5 apresenta e discute os resultados obtidos durante a execução do trabalho. Por fim, o Capítulo 6 apresenta as conclusões e contribuições desta dissertação, além de trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

Neste capítulo, apresentamos os principais trabalhos encontrados na literatura relacionados ao tema desta dissertação. O capítulo é dividido em quatro seções: a Seção 2.1 apresenta trabalhos relacionados a *Computer Assisted Language Learning* (CALL); a Seção 2.2 apresenta trabalhos que realizaram análises do Reddit; a Seção 2.3 apresenta trabalhos que estudaram a classificação automática da proficiência em idiomas. Além disso, as seções também apresentam a comparação entre os trabalhos apresentados e esta dissertação. A Seção 2.4 sumariza esta comparação, destacando o diferencial deste trabalho em relação aos outros.

2.1 Computer Assisted Language Learning

Há pelo menos duas décadas é possível encontrar trabalhos voltados para o aprendizado de línguas apoiado por tecnologias [Zhao, 1996; Warschauer & Healey, 1998]. Esse campo de estudos é chamado de *Computer Assisted Language Learning* (CALL) [Levy, 1997]. Essa área engloba quaisquer tipos de aplicações que possam auxiliar no aprendizado de uma língua estrangeira, como as redes sociais online, as quais são o foco principal desta dissertação.

O interesse dos pesquisadores de CALL em redes sociais é relativamente recente, como descrito por Zourou [2012]. Neste mesmo trabalho, a autora discute o estado da arte em relação ao uso de mídias sociais para o ensino de idiomas, mas sem especificar uma rede ou linguagem. Ela demonstra que as redes têm influência positiva no ensino e são bastante utilizadas por fomentarem a participação dos usuários.

Uma pesquisa em uma rede social específica foi apresentada pelo trabalho de Clark & Gruba [2010], que analisou a Livemocha, uma rede social (já extinta) cujo foco era o ensino de línguas estrangeiras. Neste trabalho, os autores apontam falhas

no design do modelo de interface e interação do sistema, que eram contraprodutivas sob uma perspectiva pedagógica. No mesmo trabalho os autores sugerem correções e melhorias para essas falhas na rede, com o objetivo de torná-la mais produtiva.

Arnold & Paulus [2010] analisam, sob a visão de estudantes, de um instrutor e de um observador externo, um caso prático em que uma turma real de aprendizado de certo idioma utilizou o sistema Ning, voltado para a criação de comunidades sociais. Este trabalho conclui, sob a visão do instrutor de língua, que a utilização da comunidade social para discussão teve um impacto positivo no ensino.

Sob a ótica de pessoas que estão aprendendo um novo idioma, a pesquisa realizada por Lin et al. [2016] analisa o comportamento e desenvolvimento de usuários de redes específicas para aprendizado de línguas. O estudo, feito por meio de questionários e acompanhamento de estudos de casos, conclui que as redes sociais voltadas para o aprendizado de línguas apresentam resultados positivos, mas também limitações. Além disso, conclui que para que os usuários alcancem o sucesso esperado, é necessário que as redes ofereçam apoio, orientação e atividades bem estruturadas, de maneira a promover o engajamento e interação dos usuários.

Como apresentado nesta seção, existe uma preocupação já antiga relacionada ao aprendizado de idiomas auxiliados por ferramentas computacionais, desde meados dos anos 90. Existem trabalhos na literatura voltados para a análise de softwares específicos para o apoio ao ensino de idiomas e uma linha de pesquisa específica dentro desta área tem como foco em redes sociais.

Apesar da existência de abordagens para utilização de redes sociais no auxílio do aprendizado de línguas, os trabalhos focaram em redes criadas especificamente para o escopo de ensino e aprendizado. Nesta dissertação, o objetivo é analisar comunidades específicas de aprendizado de línguas, mas que são parte de uma rede maior, de propósito geral. Isso quer dizer que essas redes de apoio ao aprendizado se formaram de maneira orgânica dentro do Reddit [Olson & Neal, 2015], ao contrário daquelas que já foram criadas com esse propósito. Por essa razão, escolhemos o Reddit como a rede a ser avaliada.

2.2 Reddit

Como o foco do presente trabalho não são redes sociais explicitamente voltadas para ensino de línguas, e sim o Reddit, é necessário compreender como as comunidades dessa rede são estruturadas analisando-se o comportamento dos usuários na rede social deste sistema. Seguindo essa linha, o trabalho de Buntain & Golbeck [2014] explora o

comportamento dos usuários, identificando que a maioria dos usuários na rede está focada em responder postagens, se voltando mais para o conteúdo e não necessariamente em interações interpessoais. No entanto, os autores não analisam especificamente redes voltadas para o aprendizado de idiomas, onde é plausível que usuários fiquem mais confortáveis em compartilhar dúvidas com um conjunto seleto de relacionamentos [Gagnon, 2013; Chester & O'Hara, 2007].

Choi et al. [2015] caracterizam padrões de conversação no Reddit, confirmando as observações de Buntain & Golbeck [2014], e demonstram que as respostas em tópicos com muitas publicações geralmente são postadas por um grupo pequeno de usuários. Além disso, a pesquisa verificou que subreddits voltados para compartilhamento exclusivo de imagens ou discussão de um tópico específico apresentam um engajamento maior dos seus usuários quando comparado a outras comunidades.

Mills [2015] apresenta uma visão geral sobre todos os subreddits existentes, explicitando que 2% destes correspondem a 98% de toda a atividade de comentários na rede. Mills categoriza essas comunidades em três grupos: aquelas com um alto número de votações diárias (voltados para criação de conteúdo), aquelas com um número próximo de votações e comentários, que são voltados para discussões de tópicos específicos, e um grupo menor, em que o número de comentários é muito maior que o número de votações. Essa categorização pode ser utilizada para auxiliar na compreensão quando analisando as redes voltadas para ensino de linguagens.

Outro trabalho envolvendo o Reddit e compartilhamento de conhecimento analisa comunidades voltadas para a área de *User Experience* (UX) [Kou et al., 2018]. Este trabalho demonstra como essas comunidades virtuais são altamente ativas, demonstrando inclusive como se dá a participação de usuários. Além disso, Weninger et al. [2013] apresenta um estudo da hierarquia de publicações no Reddit e discute maneiras com as quais o site pode ser utilizado para outros fins que não sejam necessariamente compartilhamento de notícias, que era a proposta inicial do site.

Portanto, é possível perceber que o Reddit é uma rede bastante atrativa para usuários que desejam compartilhar conteúdo, sem um foco direto e específico nas relações pessoais entre usuários. Por ser uma rede social para temas gerais, é bastante conhecida pelo grande público, o que impacta diretamente na escolha do Reddit como rede de interesse desta dissertação. Este fato é corroborado por uma pesquisa realizada no ano de 2013, que indica que 6% da população de adultos que utilizam a internet são usuários do Reddit [Duggan & Smith, 2013].

Dessa maneira, o Reddit se mostra uma rede bastante difundida e que permite a análise de um grande volume de dados, uma vez que possui um número muito maior de usuários que outras redes com focos específicos. Essa popularidade pode ser explicada

por um outro fator atrativo do Reddit é que os usuários podem se manter anônimos, o que influencia em sua desinibição e motiva a participação [Gagnon, 2013; Chester & O'Hara, 2007]. Adicionalmente, o Reddit também permite que usuários acompanhem as publicações sem precisar ter um perfil ou mesmo publicar em subreddits. Estes usuários são conhecidos como *lurkers*, aqueles que desejam "ver mais do que mostrar", como definido em Suler [2002].

Como mostrado pelo trabalho de Singer et al. [2014], o Reddit se tornou, nos últimos anos, uma rede com uma enorme variedade de conteúdo distribuído pelos seus subreddits, o que explica a existência de tópicos bem específicos, como os de aprendizado de línguas estrangeiras. Adicionalmente, o Reddit é uma rede social online aberta e seus dados estão disponíveis gratuitamente para download. Embora já existam na literatura análises de outros subreddits e principalmente análises gerais do Reddit, apresentando resultados e contribuições relevantes, este é o primeiro trabalho a investigar comunidades voltadas para o ensino de idiomas.

Em suma, nesta dissertação analisamos as relações entre os usuários dessas comunidades específicas (assim como outros pesquisadores o fizeram para subreddits de outros escopos), além do comportamento dos usuários e o conteúdo compartilhado. Por fim, investigamos como a proficiência de usuários em uma língua evolui à medida que usuários permanecem ativos na rede. Esta análise da proficiência culminou em um modelo que permite acompanhar e prever como o usuário evolui baseado nas características dos textos que publica. A Tabela 2.1 sumariza a comparação entre os principais trabalhos envolvendo o Reddit e esta dissertação.

	Objeto de análise		
	Reddit como um todo	Tipo de subreddit específico	Conteúdo dos subreddits
Buntain & Golbeck [2014]	x		
Choi et al. [2015]	x		x
Mills [2015]	x		x
Kou et al. [2018]		x (UX)	x
Weninger et al. [2013]	x		x
Gagnon [2013]	x		
Singer et al. [2014]	x		x
Este trabalho		x (Idiomas)	x

Tabela 2.1: Comparação entre esta dissertação e outros trabalhos relacionados em relação ao objeto de análise

2.3 Classificação automática da proficiência

A classificação automática da proficiência em idiomas a partir de textos também é um tema recorrente na literatura. Yang et al. [2016] demonstram que características textuais podem ser utilizadas de maneira efetiva para classificação automática. Esse trabalho utiliza resultados de testes de inglês dos mais diversos tipos, como escrita e leitura, para medir a aptidão cognitiva em um segundo idioma de um grupo de voluntários. Os resultados desses testes são utilizados para alimentar modelos de aprendizado de máquina: *Multi-Layer Perceptron*, *Naive Bayes*, Regressão Logística e *Random Forest*. Esses modelos conseguiram classificar a proficiência dos voluntários com precisão superior a 70%. Dentre os modelos testados, o *Random Forest* geralmente apresentou melhor desempenho.

Seguindo essa linha, a pesquisa de Crossley et al. [2012] realiza um trabalho similar de análise de proficiência em inglês, porém utilizando apenas características extraídas de textos escritos como dados de entrada. Esses textos foram primeiramente classificados a partir da proficiência de seus autores, baseado nos resultados de testes de proficiência, como o TOEFL. Foram utilizados 100 textos dos diferentes níveis de proficiência (iniciante, intermediário, avançado). Em seguida, esses textos foram analisados com o apoio da ferramenta Coh-Metrix[1], que quantifica, dentre outros pontos, a coesão de textos em inglês. Uma função de classificação foi utilizada sobre as características quantificadas, chegando a classificar corretamente mais de 70% dos textos. Esse trabalho corrobora os resultados da pesquisa de Crossley et al. [2011], que demonstra como textos de níveis de proficiência diferentes apresentam características diferentes.

Outros trabalhos na linha de classificação automática de proficiência presentes na literatura realizam a análise de outros meios, como os trabalhos de Zechner & Bejar [2006] e Flanagan et al. [2016], que classificam a proficiência em inglês a partir de gravações de áudio. Ambos trabalhos selecionam características dos textos das gravações para classificar a proficiência, demonstrando que é possível classificar mesmo textos não escritos.

A Tabela 2.2 compara as metodologias dos trabalhos citados nessa seção com a metodologia do presente trabalho. A análise da tabela de comparação demonstra que esta dissertação tem como diferencial utilizar não apenas os textos dos usuários, como também a própria classificação de proficiência indicada pelos próprios usuários em suas publicações para o treinamento. Outro diferencial é que utiliza apenas textos escritos, ao contrário de outras pesquisas que utilizam outros tipos de *features*, como gravações de áudio, e necessitam de análises diferentes para entradas diferentes. Além disso, a

[1] http://www.cohmetrix.com

	Proficiência classificada pelo usuário	Proficiência classificada por especialistas	Proficiência classificada por testes	Tipo de entradas	Metodologia aplicável para vários idiomas	Considera histórico dos voluntários
Yang et al. [2016]			x	Texto		
Crossley et al. [2011]			x	Texto		
Crossley et al. [2012]		x		Texto		
Zechner & Bejar [2006]			x	Áudio		
[Flanagan et al. [2016]			x	Áudio		
Este trabalho	x			Texto	x	x

Tabela 2.2: Comparação entre a metodologia aplicada em projetos da literatura e este projeto em relação a classificação automática de proficiência em um idioma.

metodologia deste trabalho pode ser replicada para outros idiomas, ao contrário dos outros trabalhos.

2.4 Diferencial do trabalho

Esta dissertação tem como diferencial realizar uma análise sobre um grupo de subreddits ainda não estudado por pesquisadores: aqueles voltados para aprendizado de idiomas. Uma metodologia específica foi criada para avaliar a proficiência dos usuários e relacioná-la às atividades (posts e comentários) dentro das redes. Os textos dos usuários foram analisados e utilizados para classificação da proficiência, a partir de modelos existentes na literatura. A partir disso propusemos um novo modelo, que considera não apenas as características dos textos, como também a curva natural da evolução da proficiência.

Diferente de outras metodologias para classificação da proficiência encontradas na literatura, que utilizam textos de avaliações específicas para proficiência, neste trabalho são utilizados textos de usuários de uma rede social. Com o apoio da ferramenta LIWC, é possível realizar a classificação da proficiência para todos os idiomas que possuem dicionários para a ferramenta, como o inglês, português, espanhol ou alemão. Isso também é um diferencial deste trabalho em comparação a outros trabalhos da literatura, que focam na língua inglesa por esta ter suporte de mais ferramentas de análise textual.

O fato de utilizar textos de uma rede social, que tem como base a classificação auto-declarada dos usuários, também faz com que a metodologia adotada seja mais acessível, visto que não são necessários testes que requerem a participação de especia-

listas treinados para avaliação de proficiência, nem a participação de voluntários para classificar os textos, como é observado na metodologia de outros trabalhos. Por fim, e mais importante, o método de classificação proposto neste trabalho leva em consideração o histórico de proficiência de um usuário, tendo em vista que utiliza o conjunto de textos que ele já publicou na rede. Dessa maneira, é possível afirmar também que a metodologia apresentada neste trabalho permite o acompanhamento da proficiência dos usuários ao decorrer do tempo, ao contrário de outros trabalhos relacionados.

Capítulo 3

Metodologia

Este capítulo tem como objetivo descrever os métodos aplicados nesta dissertação, de maneira a alcançar os objetivos como definidos anteriormente. As principais etapas da metodologia descritas neste capítulo são: a coleta dos dados do Reddit e a extração do *dataset* (Seção 3.1); a modelagem dos dados referentes à interação dos usuários em um grafo e as métricas calculadas para a essa rede (Seção 3.2); os métodos utilizados para realizar análise de interação entre usuários de diferentes proficiências e a modelagem das *threads* (Seção 3.3); o método aplicado para extrair características textuais das publicações (Seção 3.4); e os modelos de classificação experimentados para prever a proficiência dos usuários (Seção 3.5).

3.1 Coleta do *dataset*

Para realizar a análise, coletamos os dados de todos os subreddits do Reddit, que são disponibilizados online[1]. Buscamos todos os dados públicos relacionados a *posts* e comentários desde janeiro de 2010 até dezembro de 2017. Como esses dados disponibilizados englobam o Reddit em sua totalidade, extraímos os *posts* e comentários específicos dos subreddits voltados para o aprendizado de idiomas.

Após extraídos, os dados foram inseridos em um banco de dados para facilitar seu processamento. Estudamos os seguintes subreddits: French, voltado para o aprendizado de francês, *German*, para alemão, *Spanish*, para espanhol e *EnglishLearning*, para inglês, daqui para a frente referido de maneira abreviada por *English*.

[1] http://files.pushshift.io/reddit/

3.2 Modelagem e análise da rede de interações

Primeiramente, analisamos as interações entre usuários a fim de caracterizá-las. Essa caracterização auxilia na avaliação da estrutura da rede de interações como um possível facilitador do aprendizado e da evolução do nível de proficiência dos usuários em línguas estrangeiras. Para isso, modelamos os subreddits na forma de grafo, a partir do qual calculamos algumas métricas de interesse. As subseções a seguir explicam como realizamos a modelagem e as métricas analisadas.

3.2.1 Modelagem matemática

A partir dos dados extraídos, modelamos as interações entre os usuários como conexões em um grafo, sendo que foi gerado um grafo para cada um dos subreddits. Mais especificamente, representamos a rede de interações por um grafo direcionado $G = (V, E_d)$, onde V é um conjunto de vértices que representam usuários que tenham publicado um *post* ou comentário na rede e E_d é um conjunto de arestas que representam interações entre pares de usuários em V. Essas ligações são direcionadas (se o vértice v respondeu a um *post* ou comentário de um vértice u, v aponta para u). O grafo modelado possui pesos nas arestas, onde o peso é igual ao número de interações (respostas a posts ou comentários) que ocorreram entre cada par de vértices em uma determinada direção. Para o cálculo de algumas métricas, utilizamos uma versão não-direcionada $G = (V, E)$ do grafo G_d, em que cada aresta $(u, v) \in E$ indica a existência de $(u, v) \in E_d$ ou $(v, u) \in E_d$. O peso associado a $(u, v) \in E$ é igual à soma dos pesos das arestas entre os mesmos vértices na versão direcionada.

3.2.2 Métricas de interesse

Com o objetivo de compreender o comportamento dos usuários, analisamos as seguintes métricas de centralidade [Newman, 2011], para cada subreddit coletado.

- *Closeness*: esta métrica é o recíproco da média das distâncias entre um vértice v e todos os outros vértices do grafo e é calculada por

$$\text{clo}(v) = \left(\frac{\sum_{u \in V - \{u\}} d(v, u)}{n - 1} \right)^{-1},$$

  onde $d(v, u)$ é a distância entre os vértices v e u.

- Coeficiente de clusterização: esta métrica quantifica a tendência que os nós tem de se agrupar em cliques de tamanho 3. Usuários com um valor elevado para o

coeficiente de clusterização são aqueles cujos vizinhos tem alta probabilidade de interagir. A fórmula do coeficiente de clusterização c_v do vértice v para grafos com peso [Saramäki et al., 2007] é dada por

$$c(v) = \frac{1}{|\mathcal{N}(v)|(|\mathcal{N}(v)|-1)} \sum_{u \in \mathcal{N}(v),\, z \in \mathcal{N}(v)\setminus\{u\}} (\hat{w}_{vu}\hat{w}_{vz}\hat{w}_{uz})^{1/3},$$

onde $\mathcal{N}(v)$ é o conjunto de vizinhos de v em G, $\hat{w}_{ij}$ representa o peso da aresta entre os vértices i e j normalizado pelo maior peso encontrado no grafo $\hat{w}_{ij} = w_{ij}/\max_{(k,l)\in E}\{w_{kl}\}$.

- Excentricidade: esta métrica quantifica a distância máxima de um vértice u para cada um dos outros vértices da rede e é dada por

$$\mathrm{exc}(v) = \max_{u \in \mathcal{N}(v)} \{d(v,u)\}.$$

  Redes com estruturas de comunidades onde os vértices estabelecem muitos "atalhos", isto é, conexões entre subgrupos diferentes, tendem a apresentar valores mais baixos de excentricidade.

- Grau de entrada: definimos grau de entrada de v como sendo a soma dos pesos das arestas $(u,v) \in E_d$, ou seja,

$$d_{\mathrm{in}}(v) = \sum_{(u,v)\in E_d} w_{u,v}$$

  Na rede, representa o número de vezes em que usuário responderam algum *post* ou comentário publicado pelo usuário v.

- Grau de saída: analogamente, definimos o grau de saída de v como

$$d_{\mathrm{out}}(v) = \sum_{(v,u)\in E_d} w_{v,u}$$

  Na rede, representa os usuários que tiveram seus *posts* ou comentários respondidos pelo usuário u.

3.3 Análise das threads

Além de analisar as interações entre usuários dentro do subreddit, estudamos também as interações dentro dos tópicos do subreddit. Para isso, foram consideradas as

threads, tópicos criados por um usuário a partir de um tema específico e que podem ser respondidos por qualquer usuário da rede. É importante ressaltar que usuários do Reddit podem responder não apenas aos autores originais dos tópicos, mas também a comentários de outros usuários neste mesmo tópico, resultando em uma estrutura de árvore.

A partir da modelagem das *threads* através de uma estrutura de árvore, analisamos as interações dos usuários nos subreddits selecionados. Isto permitiu analisar, por exemplo, a profundidade máxima e a largura média de cada tópico. Essa análise permite que se compreenda mais detalhadamente a atividade dos usuários, se é focada na interação entre pessoas ou na especificamente com o conteúdo.

3.4 Análise textual

Uma vez que apenas o subreddit de aprendizado de alemão incentiva aos seus usuários informarem seu nível atual de proficiência no perfil, permitindo, assim, que os mesmos sejam categorizados, a etapa de análise textual foi realizada somente para esta comunidade.

Mais especificamente, nesta etapa utilizamos apenas os usuários que indicavam seu nível de proficiência no perfil, visto que mesmo com a sugestão da comunidade nem todos indicavam seu nível. Realizamos a análise textual das publicações e para isso, utilizamos a ferramenta *Linguistic Inquiry and Word Count* (LIWC), desenvolvida para a extração de atributos linguísticos do texto. O LIWC pode analisar várias línguas diferentes, e para isso possui um dicionário específico para cada.

No caso do alemão, tal dicionário ainda era o mesmo desde a primeira versão da ferramenta (2001) e não resultava em tantas características quanto os dicionários de versões mais novas (2007 ou 2015). Ainda assim, foram extraídas 81 características para os textos estudados incluindo, por exemplo, tamanho das palavras e tamanho das frases.

Para estudar a relação entre características textuais e nível de proficiência, utilizamos como base os textos de usuários que incluíram em seu perfil a proficiência no momento da publicação. A partir da extração das características textuais, realizamos uma análise para diferenciar os grupos de usuários. Para cada nível de proficiência (iniciante, intermediário, avançado e nativo) avaliamos algumas características, de maneira a constatar se os diferentes níveis de proficiência apresentam características diferentes. O resultado positivo dessa análise possibilitou a utilização de modelos de classificação para categorizar automaticamente os textos de usuários.

3.5 Modelos de classificação

A partir das características textuais extraídas pelo LIWC, treinamos diversos classificadores clássicos da literatura de aprendizado máquina para inferir o nível de proficiência associado a uma publicação: iniciante, intermediário ou avançado. Os métodos utilizados foram todos supervisionados e, portanto, utilizamos apenas publicações que possuíam *tag* de proficiência. Para selecionar o método de maior assertividade, calculamos a métrica F_1, que é a média harmônica entre precisão e revocação [Hand & Christen, 2018], para todos os resultados de modelos aplicados.

Para selecionar o melhor conjunto de hiper-parâmetros para cada um dos classificadores, utilizamos o método *grid search* [Bergstra & Bengio, 2012]. Para cada um dos modelos, selecionamos um conjunto de valores para os hiper-parâmetros e aplicamos o *grid search*, que avaliou o desempenho de cada uma das combinações possíveis de valores de hiper-parâmetros, retornando a combinação ótima. Essa avaliação é realizada a partir de testes sobre uma amostra do próprio conjunto de dados, utilizando o método de validação cruzada, que consiste na divisão do conjunto de dados em K partes mutualmente exclusivas (neste caso, K=5), de maneira que cada uma das partes sirva como conjunto de testes para o método treinado a partir das outras K-1 partes até que todas as observações do conjunto de dados tenham feito parte do conjunto de teste exatamente uma vez. Com os hiper-parâmetros dos métodos ajustados, realizamos a etapa de execução de cada um. Utilizamos a validação cruzada mais uma vez, para avaliar a eficácia de cada um dos métodos, dessa vez utilizando K=10. Em seguida, apresentamos uma descrição resumida de cada um dos classificadores utilizados.

3.5.1 KNN

KNN ou *K-Nearest-Neighbours* é um método utilizado para classificação ou regressão [Altman, 1992]. No caso de classificação, consiste em encontrar os k vizinhos mais próximos de um elemento e classificá-lo baseado no rótulo mais comum dentre eles.

Este método é supervisionado, ou seja, consiste em inferir uma função a partir de um conjunto de treinamento contendo instâncias rotuladas. Cada observação consiste em um par de características (*features*) e seu rótulo. As *features* neste trabalho são as características resultantes da aplicação do LIWC nos textos e as classes são os níveis de proficiência (iniciante, intermediário e avançado).

3.5.2 Random Forest

Random Forest é um método de aprendizado supervisionado [Breiman, 2001] que pode ser utilizado para regressão e também classificação. Este algoritmo consiste na criação de uma série de árvores de decisão a partir de vários subconjuntos do conjunto de treino original.

Esta técnica é muito utilizada para classificação porque a utilização de muitas árvores de decisão curtas (i.e., com poucas variáveis) melhora a precisão dos resultados e reduz o *overfitting*. Durante a classificação de novas observações, são inseridas apenas *features* e o método, já treinado, classifica esse dado percorrendo a árvore de decisão.

3.5.3 Regressão Logística

A Regressão Logística é um método de classificação [Schmidt et al., 2017; Yu et al., 2011] que, a partir de um conjunto de variáveis, estima a probabilidade de um dado representado por determinadas features pertencer a uma determinada classe. Este método considera uma combinação linear das *features* mapeada para o intervalo $[0, 1]$, por meio da função logística [Freedman, 2009]. Portanto, a probabilidade de o texto assumir determinada classe Y dadas as variáveis independentes $X_1, \ldots, X_n$ é

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}},$$

onde

$$g(x) = B_0 + B_1 X_1 + ... + B_n X_n.$$

Esse método realiza uma abordagem *one-versus-all*, calculando a probabilidade de o texto assumir ou não determinada classe (proficiência), sem discriminar as outras.

3.5.4 Gradient Boosting e XGBoost

O método *Gradient Boosting* é um método de predição que cria um modelo forte a partir de um conjunto de modelos mais fracos, de maneira iterativa, como por exemplo, realizando a combinação de resultados de árvores de regressão [Friedman et al., 2001; Friedman, 2001, 2002]. Utilizamos esse modelo como um classificador da proficiência dos textos.

O método *XGBoost* é uma versão otimizada do Gradient Boosting que permite resolver problemas mais rapidamente com a assertividade próxima à do *Gradient Boosting* [Chen & Guestrin, 2016].

3.5.5 Classificação sequencial

Aplicamos todos os métodos de classificação anteriores a cada publicação rotulada para comparar suas assertividades. Utilizamos as previsões dos classificadores que apresentaram os melhores resultados como entrada para um modelo que leva a sequência de publicações de um usuário em consideração. Como esse novo modelo é uma das principais contribuições dessa pesquisa, ele será explicado detalhadamente no Capítulo 4.

Capítulo 4

SEMPLICe

Neste capítulo apresentamos um modelo probabilístico que considera o histórico de publicações na clssificação da proficiência de um usuário. O modelo é denominado de SEMPLICe (SEquential Model for Proficiency cLassifICation)[1].

4.1 Modelo SEMPLICe de evolução de proficiência

Desejamos encontrar a sequência de níveis de proficiências $\mathbf{Y} = \{Y_1 \leq Y_2 \leq \ldots \leq Y_N\}$, que maximiza a probabilidade condicional $P(Y_1, \ldots, Y_N | X_1, \ldots, X_N)$, onde X_i, $i = 1, \ldots, N$, representa o conjunto de *features* da publicação i. SEMPLICe é baseado em duas suposições. A primeira suposição é de que o nível de proficiência permanece o mesmo ou aumenta de Y_t para Y_{t+1}. Esta suposição é razoável se considerarmos usuários que não permanecem inativos durante muito tempo. A segunda suposição é de que, condicionado à restrição de que Y_t é não-decrescente, a probabilidade $P(Y_1, \ldots, Y_n | X_1, \ldots, X_N)$ pode ser aproximada por um produto de fatores:

$$P(Y_1, \ldots, Y_n | X_1, \ldots, X_N) \approx \prod_{t=1}^{N} P(Y_t | X_t), \quad \text{onde } Y_1 \leq Y_2 \leq \ldots Y_N. \tag{4.1}$$

A base lógica para a segunda suposição é de que probabilidades de transição do tipo $P(Y_{t+1}|Y_t)$ variam de usuário para usuário conforme fatores externos, não podendo ser aprendidas de forma efetiva a partir do conjunto de dados que utilizamos. Embora seja possível usar o estimador frequentista para $P(Y_{t+1} = j | Y_t = i)$ (isto é, a razão entre o número de transições $i \to j$ e o número de transições de i para qualquer estado), tal estimador representa uma média de uma população potencialmente muito diversa.

[1] Em italiano, semplice significa *simples*

O problema de se encontrar a sequência de rótulos $\mathbf{Y} = \{Y_t\}_{t=1}^{N}$ que maximiza o lado direito de (4.1) é, após a transformação logarítmica, equivalente ao problema de otimização

$$\max_{\mathbf{Y}: Y_1 \leq Y_2 \leq \ldots \leq Y_N} \sum_{i=1}^{N} \log P(Y_t|X_t). \tag{4.2}$$

As probabilidades $P(Y_t|X_t)$ são calculadas a partir dos modelos de classificação treinados na abordagem não-sequencial. Primeiro, é realizado o treinamento do modelo, que é feito a partir de um subconjunto aleatório das publicações. Então, na etapa de predição, o modelo recebe como entrada as *features* X_t de uma publicação t, e retorna um vetor $[p_b, p_i, p_a]$, que contém a probabilidade de que aquela publicação seja classificada em cada nível de proficiência, sendo p_b a probabilidade de ser iniciante, p_i, de ser intermediário e p_a, avançado.

4.1.1 Diferença em relação ao HMM

O modelo proposto se assemelha com o *Hidden Markov Model* (HMM) ou Modelo Oculto de Markov. Esse modelo estatístico é muito utilizado para modelar séries temporais formadas por uma componente não-observável e outra componente observável. No nosso trabalho, a componente não-observável corresponde ao nível de proficiência de um indivíduo, enquanto a componente observável corresponde ao vetor de features textuais extraído das publicações.

No HMM, a componente não-observável é uma sequência de estados. Estes estados são ocultos pois não é possível determinar exatamente em que o estado o processo se encontra a cada passo. As transições entre os estados são controlados por uma cadeia de Markov de tempo discreto. Por outro lado, a componente observável é uma sequência de símbolos emitidos pelo processo. A probabilidade de emissão de cada símbolo em um determinado passo depende do estado oculto em que o processo se encontra.

Seria possível forçar uma correspondência dos estados ocultos a níveis de proficiência definindo-se a cadeia que governa a transição desses estados como um grafo linha formado por três estados, conectados da esquerda para direita, com *self-loops* em cada estado. Contudo, seria complicado definir uma distribuição de símbolos, dado um determinado estado. Isso acontece porque o tamanho do conjunto de símbolos é exponencial no número de features e porque a grande maioria dos símbolos não são observados no conjunto de dados. Ao invés disso, o que o SEMPLICe faz é trabalhar com uma distribuição sobre os estados, dado o símbolo que foi emitido. Esta distribuição é obtida a partir de modelos de classificação baseados em aprendizado supervisionado.

Para resolver o problema de otimização em (4.2) é proposto um algoritmo de

Algoritmo 1: Algoritmo sequencial: versão força bruta

entrada: matriz $\mathbf{M}_{3\times N}$, onde $M[i,t] = \log P(Y_t = i|X_t)$
saída : par TRANSICOES $= (t_i, t_a)$ indicando que a transição para intermediário acontece no passo t_i e para avançado, no passo t_a

```
/* Caso em que não há transições                                    */
1  TRANSICOES ← (Nunca, Nunca)
2  MAXLOGLIKELIHOOD ← ∑_{t=1}^{N} M[1,t]
3  for t_i de 1 até N do
       /* Caso em que se torna intermediate em t_i                  */
4      LOGLIKELIHOOD ← ∑_{t=1}^{t_i-1} M[1,t] + ∑_{t=t_i}^{N} M[2,t]
5      if LOGLIKELIHOOD > MAXLOGLIKELIHOOD then
6          MAXLOGLIKELIHOOD ← LOGLIKELIHOOD
7          TRANSICOES ← (t_i, Nunca)
8      for t_a de t_i + 1 até N do
           /* Caso em que transições ocorrem em t_i e t_a           */
9          LOGLIKELIHOOD ← ∑_{t=1}^{t_i-1} M[1,t] + ∑_{t=t_i}^{t_a-1} M[2,t] + ∑_{t=t_a}^{N} M[3,t]
10         if LOGLIKELIHOOD > MAXLOGLIKELIHOOD then
11             MAXLOGLIKELIHOOD ← LOGLIKELIHOOD
12             TRANSICOES ← (t_i, t_a)
13 return TRANSICOES
```

programação dinâmica na Seção 4.3. No entanto, para um melhor entendimento, na Seção 4.2 será descrito primeiro um algoritmo de força bruta, cuja complexidade é $\Theta(N^2)$, dada sua implementação baseada em dois laços aninhados.

4.2 Algoritmo I: Força Bruta

O Algoritmo 1 descreve em pseudocódigo a versão força bruta do algoritmo sequencial. Para estimar a sequência de rótulos para um usuário com N publicações, o algoritmo recebe como entrada uma matriz $\mathbf{M}_{3\times N}$, onde $M[i,t] = \log P(Y_{i,t}|X_t)$, onde $i \in \{1,2,3\}$ representam os níveis de proficiência *beginner*, *intermediate* e *advanced*, respectivamente. A saída do algoritmo é um par TRANSICOES $= (t_i, t_a)$ indicando que o usuário passou a ter nível intermediário em t_i e avançado em t_a.

O algoritmo calcula o log-likelihood da sequência para todas as combinações possíveis de instantes de transição (de iniciante para intermediário e de intermediário para avançado). Isso é possível utilizando dois laços aninhados, onde o primeiro (t_i), define a publicação a partir da qual o usuário passa a ser intermediário (partindo de 1, i.e., nenhuma publicação iniciante). O segundo ($t_a > t_i$) define qual a primeira

publicação em que o nível de proficiência do usuário é avançado.

Entre as linhas 4 e 7, considera-se o caso em que o usuário permanece no nível 2 (intermediário) até o fim da sequência. Entre as linhas 8 e 12, consideramos o caso em que fazer uma outra transição para o nível 3 (avançado) é mais provável. A complexidade computacional do algoritmo é $\Theta(N^2)$. Dentre as possíveis sequências, três apresentam apenas uma proficiência, $2 \times (N-1)$ apresentam apenas uma transição e $(N-1)(N-2)/2$ apresentam duas.

4.2.1 Exemplo de execução

As Figuras 4.1 a 4.3 apresentam um exemplo da execução da versão força-bruta do algoritmo do SEMPLICe. Na Figura 4.1, apresentamos a matriz M, com os logaritmos das probabilidades de cada publicação apresentar um nível de proficiência.

	P_1	P_2	P_3	P_4	P_5	P_N
Iniciante	-0.3	-0.5	-0.3	-0.4	-0.7	-0.5
Intermediário	-0.7	-0.3	-0.5	-0.4	-0.5	-0.4
Avançado	-0.5	-0.7	-0.7	-0.7	-0.3	-0.5

Figura 4.1: Matriz M

A versão força-bruta do SEMPLICe analisa todas as possibilidades de transição entre os níveis e calcula o somatório das probabilidades. Na Figura 4.2, apresentamos alguns exemplos, com transições em diferentes momentos da atividade do usuário. O algoritmo retorna aquele caminho composto pelos dois momentos de transição em que o somatório dos logaritmos das probabilidades tem maior valor, como mostramos na Figura 4.3.

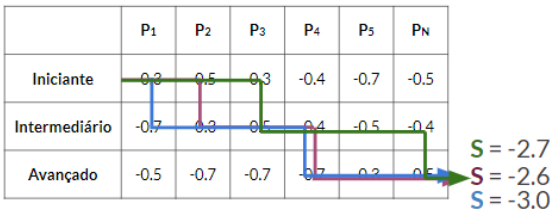


Figura 4.2: Exemplos de momentos de transição de nível de proficiência.

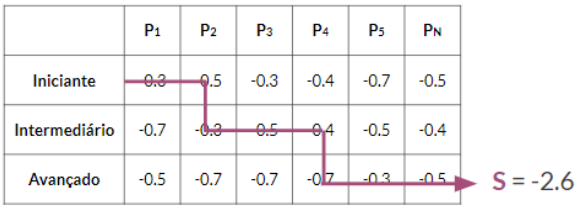


Figura 4.3: Caminho com maior somatório dos logaritmos das probabilidades retornado pelo algoritmo.

4.3 Algoritmo II: Programação dinâmica

O Algoritmo 2 descreve a versão do SEMPLICe baseada em programação dinâmica, que tem as mesmas entradas e saídas da versão força bruta, mas cuja complexidade é $\Theta(N)$ em vez de $\Theta(N^2)$. Esta versão se baseia na construção de uma memória de cálculo representada pela matriz $\mathbf{R}$, onde

$$\mathbf{R}[p,i] = \max_{p \leq Y_i \leq Y_{i+1} \leq \ldots \leq Y_N} \sum_{t=i}^{N} \log P(Y_t|X_t),$$

ou seja, o máximo da verossimilhança da subsequência $\{Y_i, \ldots, Y_N\}$, dado que $Y_i \geq p$.

Para calcular $\mathbf{R}[p,t]$, é necessário determinar se é mais provável transicionar para um nível de proficiência superior ou permanecer no nível p no instante t. No primeiro caso, o máximo da função de log-verossimilhança ao transicionar para $p+1$ imediatamente antes de fazer a publicação X_t é dado por $\mathbf{R}[p+1,t]$. No segundo caso, a máximo da função de log-verossimilhança é dado pela soma de $\mathbf{M}[p,t] = \log P(Y_t = p|X_t)$ e $\mathbf{R}[p,t+1]$. Portanto, para $p \in \{1,2\}$ e $t \in \{1,\ldots,N-1\}$, tendo a relação de recorrência

$$\mathbf{R}[p,t] = \max\left(\mathbf{R}[p+1,t],\, \mathbf{M}[p,t] + \mathbf{R}[p,t+1]\right). \tag{4.3}$$

A partir da Eq. (4.3) observa-se que a matriz $\mathbf{R}$ pode ser construída de "baixo para cima" e, dentro de cada linha, da "esquerda para a direita". Falta, portanto, definir como a última linha e a última coluna serão inicializadas. Tendo em vista que o nível máximo de proficiência é 3 (avançado), inicializamos $\mathbf{R}[3,i] = \sum_{t=i}^{N} \mathbf{M}[3,t]$ entre as linhas 1 e 3 do pseudocódigo, ou seja, considerando que não haverá novas transições. Já para preencher a coluna $\mathbf{R}[p,N]$, fazemos

$$\mathbf{R}[p,N] = \max\left(\mathbf{R}[p+1,N],\, \mathbf{M}[p,N]\right). \tag{4.4}$$

Este algoritmo retorna exatamente o mesmo resultado que a versão força bruta.

Algoritmo 2: Algoritmo sequencial: versão programação dinâmica

entrada: matriz $\mathbf{M}_{3\times N}$, onde $M[i,t] = \log P(Y_t = i|X_t)$
saída : par TRANSICOES $= (t_i, t_a)$ indicando que a transição para intermediário acontece no passo t_i e para avançado, no passo t_a

```
/* Inicialização da última linha de R                              */
R[3, N] ← M[3, N]
for t de N − 1 até 1 do
    R[3, t] = M[3, t] + R[3, t + 1]
TRANSICOES ← (Nunca, Nunca)
for p de 2 até 1 do
    for t de N até 1 do
        if t = N then
            /* Máximo, caso termine sequência em p                 */
            LOGLIKELIHOOD ← M[p, t]
        else
            /* Máximo, caso proficiência seja p em t               */
            LOGLIKELIHOOD ← M[p, t] + R[p, t + 1]
        if R[p + 1, t] > LOGLIKELIHOOD then
            /* Trocar para p + 1 em t é mais provável              */
            R[p, t] ← R[p + 1, t]
            TRANSICOES[p] = t
        else
            /* Permanecer em p é mais provável                     */
            R[p, t] ← LOGLIKELIHOOD
return TRANSICOES
```

No entanto, a análise de complexidade desta variante demonstra que ela é $\Theta(N)$, representando uma redução significativa em relação à versão anterior, que é $\Theta(N^2)$, especialmente para sequências longas. A variante baseada em programação dinâmica é a que foi de fato utilizada para a modelagem dos dados reais.

4.3.1 Exemplo de execução

As Figuras 4.4 a 4.8 exemplificam a execução da versão em programação dinâmica do algoritmo do SEMPLICe. Por conveniência, mostramos novamente a matriz M na Figura 4.4, enquanto a Figura 4.5 apresenta a matriz R, com a última linha já calculada.

O laço principal do algoritmo se inicia pelo cálculo da última célula da penúltima linha e segue caminhando para as células à esquerda e, em seguida, para a linha acima, avaliando o valor de cada célula. A Figura 4.6 mostra como o algoritmo trata a última

	P1	P2	P3	P4	P5	PN
Iniciante	-0.3	-0.5	-0.3	-0.4	-0.7	-0.5
Intermediário	-0.7	-0.3	-0.5	-0.4	-0.5	-0.4
Avançado	-0.5	-0.7	-0.7	-0.7	-0.3	-0.5

Figura 4.4: Matriz M

	P1	P2	P3	P4	P5	P6
Iniciante	-0.3	-0.5	-0.3	-0.4	-0.7	-0.5
Intermediário	-0.7	-0.3	-0.5	-0.4	-0.5	-0.4
Avançado	-3.4	-2.9	-2.2	-1.5	-0.8	-0.5

Figura 4.5: Matriz R com a última linha inicializada

	P1	P2	P3	P4	P5	P6
Iniciante	-0.3	-0.5	-0.3	-0.4	-0.7	-0.5
Intermediário	-0.7	-0.3	-0.5	-0.4	-0.5	-0.4
Avançado	-3.4	-2.9	-2.2	-1.5	-0.8	-0.5

P = max(-0.4, -0.5)

Figura 4.6: Exemplo de tratamento da última célula da linha da Matriz R

	P1	P2	P3	P4	P5	P6
Iniciante	-0.3	-0.5	-0.3	-0.4	-0.7	-0.5
Intermediário	-0.7	-0.3	-0.5	-0.4	-0.5	-0.4
Avançado	-3.4	-2.9	-2.2	-1.5	-0.8	-0.5

P = max(-0.9, -0.8)

Figura 4.7: Exemplo de tratamento da penúltima célula da linha da Matriz R

célula das linhas: compara o valor da célula atual com o da célula abaixo e atualiza o valor com aquele de maior probabilidade.

	P1	P2	P3	P4	P5	P6
Iniciante	-0.3	-0.5	-0.3	-0.4	-0.7	-0.5
Intermediário	-0.7	-0.3	-0.5	-0.4	-0.8	-0.4
Avançado	-3.4	-2.9	-2.2	-1.5	-0.8	-0.5

Figura 4.8: Matriz R com valor da penúltima célula da última linha atualizado e indicando transição de nível

A Figura 4.7 apresenta como o algoritmo trata as células que não estão na última coluna: soma seu valor ao valor da célula à direita (no caso, totalizando -0.9) e compara com o valor da célula abaixo (no caso, -0.8). Caso a célula abaixo apresente uma probabilidade maior, como no exemplo, este valor é copiado para a célula atual, como mostrado na Figura 4.8, e a coluna atual então indica o momento em que ocorreu a transição da proficiência atual para o próximo nível.

Capítulo 5

Resultados

Este capítulo apresenta os resultados obtidos a partir da execução da metodologia descrita no Capítulo 3. Cada seção contém os resultados da execução da respectiva etapa da metodologia. Apresentamos, os resultados da coleta do *dataset* e seleção dos subreddits (seção 5.1); da modelagem e análise das redes de interações (seção 5.2); da análise das *threads* iniciadas dentro do subreddit e das interações entre usuários (seção 5.3); da análise textual das publicações dos usuários, comparando as características de publicações de diferentes níveis de proficiência (seção 5.4); da aplicação dos métodos de aprendizado de máquina para classificação do nível de proficiência dos usuários utilizando as características dos textos como insumo (seção 5.5); e da aplicação do SEMPLICe, que considera não apenas as características textuais, mas também a ordem das publicações (seção 5.6). Além disso, essa seção compara o desempenho deste método com o desempenho de classificadores tradicionais.

5.1 Coleta do dataset e seleção das redes analisadas

A primeira parte da metodologia consistiu na análise das redes de interação nas comunidades do Reddit. Selecionamos os subreddits voltados para o aprendizado de alemão, espanhol, francês e inglês, entre janeiro de 2010 e dezembro de 2017. Os dados coletados foram filtrados para manter apenas usuários únicos e cujo perfil ainda existia na comunidade. Após essa filtragem inicial, modelamos analisamos e comparamos as redes de interação de cada subreddit. A Tabela 5.1 apresenta os dados sumarizados de cada subreddit após a modelagem no grafo G_d, incluindo o número de usuários, ligações entre usuários, *posts* e de comentários de cada comunidade.

	Número de usuários	Número de ligações	Número de *posts*	Número de comentários
English	3736	8050	5493	11967
French	18113	95152	17329	145700
German	17808	98150	18059	141346
Spanish	13648	74108	14295	113843

Tabela 5.1: Quantidade de usuários, ligações, posts e comentários de cada subreddit

5.2 Modelagem e análise das redes de interações

A partir do grafo direcionado com pesos representando o número de interações no grafo direcionado G_d, calculamos a distribuição conjunta de graus de entrada e de saída dos vértices. O grau de entrada de um vértice é a quantidade de comentários que o usuário correspondente recebeu em suas publicações. Por outro lado, o grau de saída de um vértice é a quantidade de publicações feitas por um usuário. A distribuição conjunta é mostrada na Figura 5.1 através de um mapa de calor em escala log-log, onde a cor do ponto (i, j) indica a fração de vértices em G_d com grau de entrada $i - 1$ e grau de saída $j - 1$.

Para todos os subreddits, observamos uma distribuição de cauda pesada em relação a ambas as distribuições marginais, além de uma forte correlação entre grau de entrada e grau de saída. Em sua grande maioria, usuários fazem e recebem poucos comentários, enquanto alguns poucos que fazem muitos comentários também recebem muitas respostas em suas publicações. Em particular, o subreddit English se destaca por apresentar uma quantidade muito pequena de nós com graus de entrada ou saída maiores que 10^2, o que pode ser explicado por ter menos usuários que outras comunidades.

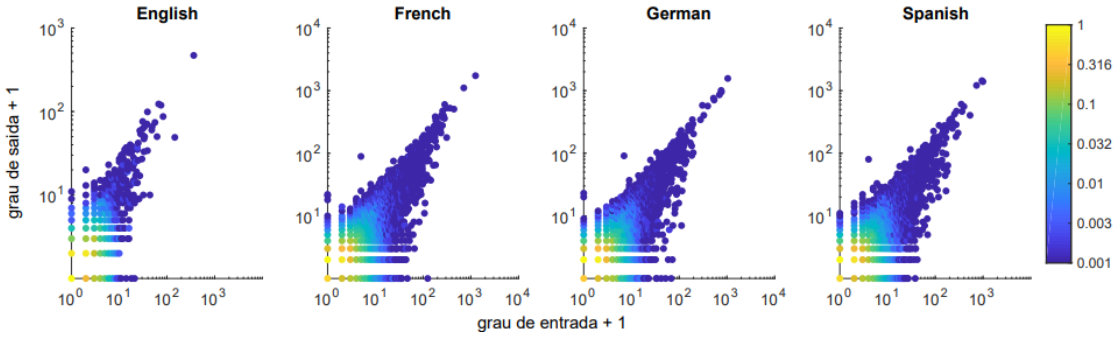


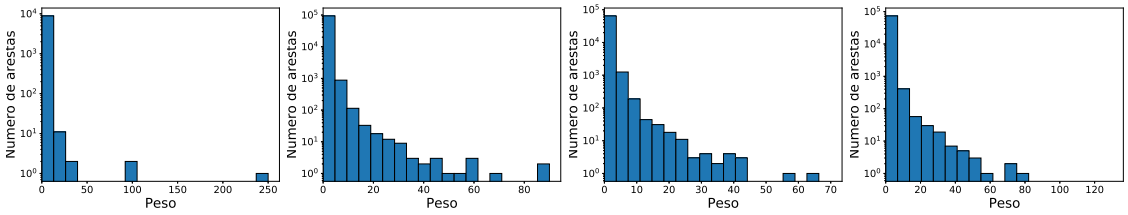
Figura 5.1: Distribuição conjunta do grau de entrada e saída. Cor indica a fração de participantes com dado grau de entrada e saída.

Embora a distribuição de graus de entrada e saída em um subreddit tenham cauda

pesada assim como grande parte das redes sociais online, é natural ponderar se os grafos G_d e G exibem características locais semelhantes a essas redes. Para responder esta questão, mensuramos algumas destas características através das seguintes métricas:

- Pesos das arestas: medem a intensidade da interação entre pares de usuários. Definido como número de interações entre um par durante o intervalo considerado.
- Centralidade de closeness: mede o quão próximo um vértice está dos outros. Definido como o inverso da soma das distâncias entre um vértice e cada um dos outros vértices alcançáveis.
- Coeficiente de clusterização: mede o quão conectados estão os vizinhos de um nó. Definido como a fração de arestas existentes entre vizinhos.

A Figura 5.2 mostra os histogramas de pesos nas arestas obtidos para cada subreddit. É possível observar que a maioria arestas tem peso 1, indicando apenas uma interação entre dois usuários. Poucas arestas possuem peso elevado (p. ex., acima de 20), o que indica que poucos são os pares de vértices com muitas interações.

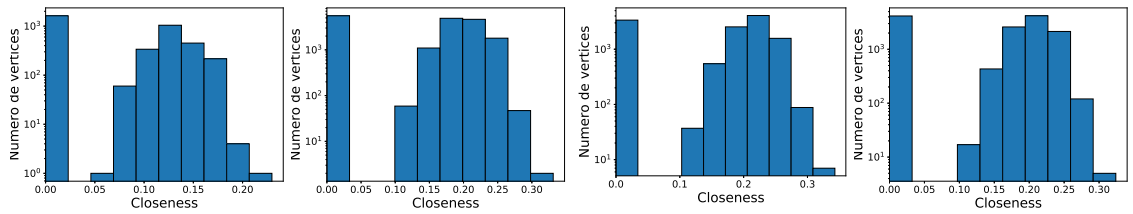


(a) English (b) French (c) German (d) Spanish

Figura 5.2: Distribuição do peso das arestas.

A Figura 5.3 mostra os histogramas da distribuição do *closeness* para cada subreddit. Inspecionando os valores que compõem a barra mais à esquerda de cada histograma, descobrimos que correspondem a vértices cujo closeness é exatamente zero. Estes são vértices isolados que fizeram uma ou mais publicações e que não obtiveram resposta. Observam-se valores baixos, indicando distâncias longas entre vértices. Isto corrobora a intuição de que os usuários não têm interesse específico em conhecer pessoas, e sim em compartilhar conteúdo de interesse à rede. As conexões são formadas conforme os tópicos de interesse comum são compartilhados.

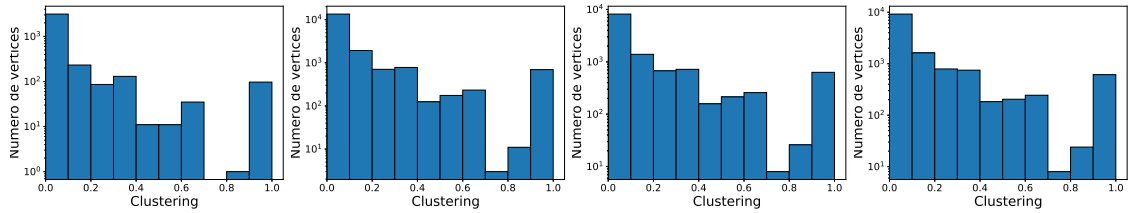
A Figura 5.4 apresenta os histogramas do coeficiente de clusterização nas redes. É possível observar que um número elevado de usuários apresenta valores baixos para esta métrica, o que indica que as redes são esparsas e não apresentam muita formação de triângulos, como no caso de redes sociais tradicionais. Isso reflete novamente na



(a) English (b) French (c) German (d) Spanish

Figura 5.3: Distribuição do *closeness*.

principal característica do Reddit, que é centrado em conteúdo em vez de amizade entre usuários.



(a) English (b) French (c) German (d) Spanish

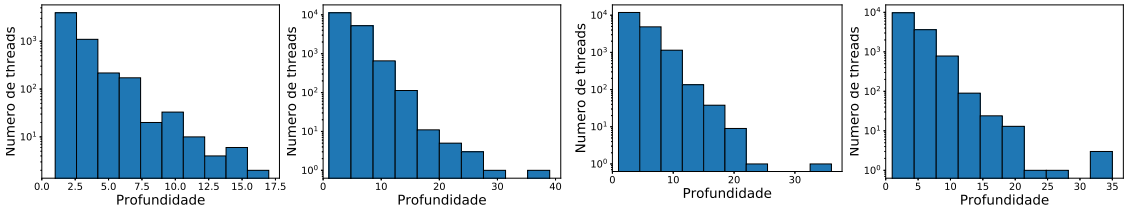
Figura 5.4: Distribuição do coeficiente de clusterização.

5.3 Análise das *threads*

Além da análise dos grafos, também analisamos os *posts* e comentários dos subreddits através das árvores de discussão. Primeiramente, investigamos a distribuição da profundidade destas árvores. A profundidade é um indicador da progressão de discussões, já que uma árvore com muitos níveis indica que na *thread* correspondente houve pelo menos uma longa cadeia de comentários.

A Figura 5.5 mostra a distribuição das profundidades para cada um dos subreddits. Observa-se que a maioria das postagens dão origem a árvores com baixa profundidade e que são poucas as postagens que terminam em discussões longas. As árvores de profundidade 1 são aquelas cujos *posts* não receberam nenhum comentário. Por conveniência, mostramos a fração de *posts* respondidos em cada subreddit na Tabela 5.2. Os subreddits French, German e Spanish apresentam altos índices de respostas às dúvidas compartilhadas pelos membros.

Para alcançar uma melhor compreensão do que leva uma *thread* a se prolongar por muitos níveis, realizamos uma análise qualitativa daquelas com a maior profundidade



(a) English (b) French (c) German (d) Spanish

Figura 5.5: Distribuição da profundidade das árvores de discussão.

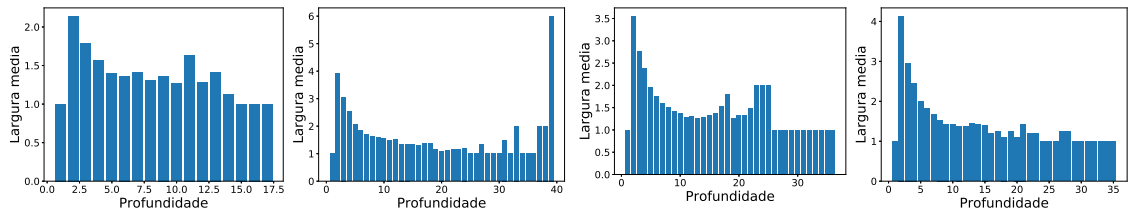
	English	**French**	**German**	**Spanish**
Com resposta	53,55%	83,88%	88,15%	81,56%
Sem resposta	46,45%	16,12%	11,85%	18,44%

Tabela 5.2: Porcentagem de *posts* com e sem respostas.

considerando cada subreddit. Constatamos que no subreddit *German*, a árvore mais profunda é voltada ao compartilhamento de dicas de pronúncia, enquanto no *French* o tema da árvore correspondente está relacionado a dicas para melhora do vocabulário. Para o *English*, a *thread* de maior profundidade discute as diferenças entre o inglês coloquial e forma culta. Na comunidade *Spanish* a árvore de discussão mais longa tem um tema menos voltado para proficiência e mais para a vivência, dado que a discussão gira em torno de como a imersão na cultura estrangeira é uma das melhores maneiras de se aprender um novo idioma. Esta análise provê indícios de que *threads* de maior engajamento dos usuários tendem a ter como assunto principal conselhos para melhora da proficiência de um usuário.

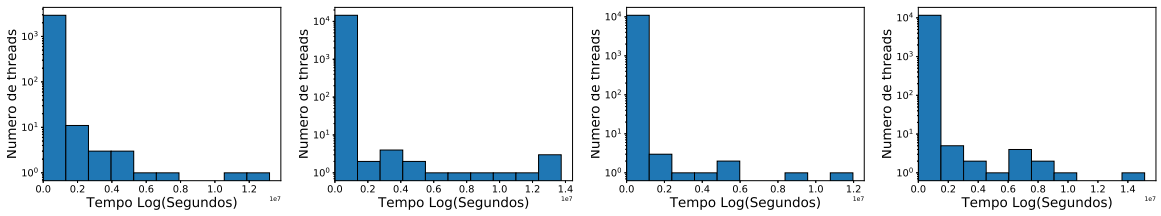
Em seguida, investigamos a largura média das árvores da discussão por nível da árvore, ou seja, a quantidade média de publicações que cada *thread* teve em cada nível de profundidade. A Figura 5.6 mostra a distribuição média da largura por nível de profundidade nos subreddits. Como o primeiro nível da árvore contém apenas o post que dá início a thread, a largura de cada árvore no nível 1 é exatamente 1. É possível observar que o engajamento dos usuários é muito elevado nos primeiros níveis de profundidade, ou seja, os usuários respondem diretamente ao post ou aos primeiros comentários. Isso corrobora a ideia de que as discussões não se prolongam por muitos comentários e que são poucas as *threads* em que a interação dos usuários ocorre por muito tempo.

Outra métrica importante de engajamento dos usuários é o tempo decorrido até que um post seja respondido pela primeira vez. A Figura 5.7 mostra o tempo em segundos entre o momento em que o *post* é publicado e o momento em que recebe o primeiro



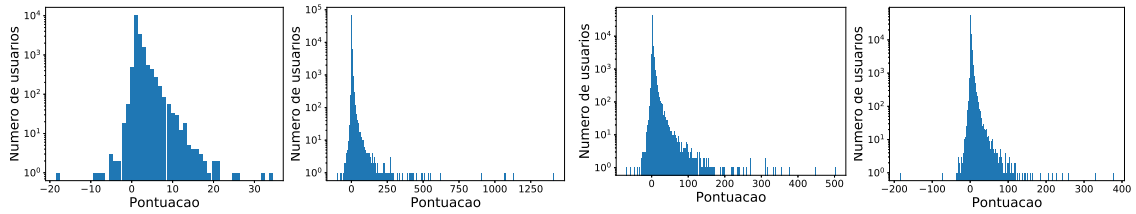
(a) English (b) French (c) German (d) Spanish

Figura 5.6: Largura média por nível das árvores de discussão (condicionado em profundidade > nível).



(a) English (b) French (c) German (d) Spanish

Figura 5.7: Distribuição do tempo decorrido até a primeira resposta em uma *thread*.



(a) English (b) French (c) German (d) Spanish

Figura 5.8: Distribuição da pontuação das *threads*.

comentário. Observamos que, dentre os *posts* que receberam uma resposta, a grande maioria foi respondida rapidamente e que poucos demoraram até serem respondidos. Esse fato mais uma vez corrobora com a intuição de que os usuários se voltam para o conteúdo, dado que as respostas são rápidas e raras vezes as *threads* se prolongam.

Uma outra métrica que avaliamos está relacionada à pontuação dos *posts*, calculada pela diferença entre o número de *upvotes* e *downvotes*. O próprio Reddit ordena as publicações pela pontuação final: quanto maior, mais destaque tem o *post* e, consequentemente, maior a probabilidade de que usuários venham a interagir nesta *thread*.

A Figura 5.8 mostra a distribuição da pontuação dos *posts* por subreddit. É possível observar que a pontuação possui valores baixos, em torno de zero. Isso quer dizer que os usuários não têm muito costume de votar nos *posts*, o que reforça a ideia

de que as interações são voltadas para o conteúdo e não aprofundam nas discussões.

5.4 Análise textual

Conhecendo melhor o comportamento dos usuários, analisamos então os textos de suas publicações. Para essa análise, utilizamos apenas o subreddit German, pois este solicita explicitamente aos usuários que adicionem *tags* (*flairs*) de proficiência a seus perfis. Consequentemente, o *German* possui um número muito grande de usuários que indicam seu nível de fluência quando comparado às outras comunidades. Essa comunidade possui um total de 159.407 publicações entre janeiro de 2010 e dezembro de 2017.

Do total de publicações do subreddit German neste período, o número daquelas com *tags* de proficiência Iniciante, Intermediário e Avançado é, respectivamente, 9.569, 12.211 e 8.026. A partir dos *posts* associados a usuários com certo nível de proficiência, realizamos uma análise textual preliminar usando a ferramenta LIWC. Avaliamos quatro propriedades: (i) o número de palavras no texto; (ii), o número de palavras por frase; (iii) a quantidade de palavras com mais de seis letras; e (iv) a quantidade de palavras reconhecidas pelo dicionário do LIWC.

A Tabela 5.3 mostra as médias e medianas de cada propriedade, para cada grupo de publicações. Os valores mais elevados para cada propriedade aparecem em negrito. Observa-se que usuários com alemão avançado ou de língua nativa tendem a usar mais palavras nos *posts*. Além disso, estes usuários tendem a usar mais palavras por frase. Considerando "Nativo" como mais proficiente que "Avançado", observa-se que o tamanho das palavras utilizadas cresce com a proficiência e que o uso de palavras do dicionário diminui com ela (possivelmente dando lugar a gírias e expressões idiomáticas).

		Iniciante	**Intermediário**	**Avançado**	**Nativo**
Contagem de palavras	Mediana	17,00	18,00	**29,00**	25,00
	Média	44,36	40.44	54,12	**58,77**
Palavras por frase	Mediana	7,00	8,30	**11,00**	10,25
	Média	7,67	8,98	**11,63**	11,44
Palavras grandes	Mediana	16,67	19,35	21,01	**21,05**
	Média	17,29	19,61	21,25	**22,39**
Palavras do dicionário	Mediana	**64,71**	64,58	62,96	59,04
	Média	**66,32**	65,46	63,27	60,67

Tabela 5.3: Resultados da análise textual básica utilizando a ferramenta LIWC.

Os resultados da análise textual das publicações indicam que existem diferenças

fundamentais na forma como as pessoas em diferentes níveis de proficiência escrevem no subreddit. Esse fato pode ser utilizado para estimar a proficiência do usuário, quando esta é desconhecida. Contudo, esta análise possui duas limitações: a proficiência é auto-declarada e, portanto, as médias e medianas podem estar super- ou subestimadas; apesar da comunidade solicitar o uso das *tags* explicitamente, 80% dos usuários não as tem em seus perfis, podendo o viés daqueles que as tem ser significativo sobre as métricas estudadas.

5.5 Modelos de classificação

Após a análise dos textos e geração de características, demos início a etapa de classificação da proficiência a partir de métodos de aprendizado de máquina encontrados na literatura. Utilizamos cinco métodos: Logistic Regression, Random Forest, KNN e Gradient Boosting e XGBoost. Todos foram implementados utilizando a linguagem Python e a biblioteca Scikit. Os métodos utilizados permitem a execução *multiclass*, já que os textos foram classificados em três classes (iniciante, intermediário e avançado).

Para reduzir o tamanho do conjunto de *features* utilizado, aplicamos o método *chi-squared* que, como apontado por Spolaôr & Tsoumakas [2013], é um excelente método para seleção de *features* para classificação textual. Este método seleciona as K melhores *features* do conjunto, sendo K um valor que variamos entre 5 e 80. Os resultados utilizando menos *features* que o número máximo apresentaram uma precisão mais baixa e, por isso, utilizamos todas as *features* disponibilizadas.

Cada um dos modelos de classificação possui hiper-parâmetros que precisam ser escolhidos antes que seja feito o ajuste aos dados. Para selecionar o valor de cada um dos hiper-parâmetros, utilizamos o método *grid search* [Bergstra & Bengio, 2012]. Este método indica para cada modelo quais parâmetros deseja-se avaliar, e realiza testes para todas as combinações possíveis. A Tabela 5.4 mostra os parâmetros e respectivos valores testados para cada modelo. Os valores selecionados pela busca estão destacados em negrito.

Após a definir e fixar os valores dos hiper-parâmetros, avaliamos o desempenho de cada um dos classificadores. A metodologia escolhida para treinamento e teste dos métodos foi a validação cruzada utilizando *K-fold* [Kohavi et al., 1995]. Neste método, o conjunto de dados é dividido em K partes iguais e o modelo é testado K vezes para cada uma dessas partes, sendo as outras $K - 1$ utilizadas para o treinamento. Para cada classificador, então, calculamos a média e o desvio padrão da métrica F1 dos resultados, apresentados na Tabela 5.5.

KNN	
Hiper-Parâmetro	**Valores**
K	3, 5, 10, 20, 50, 100, **200**, 500
Tam. folhas	2, **5**, 10
Random Forest	
Hiper-Parâmetro	**Valores**
Núm. Estimadores n	32, 64, **96**, 128
Máx. Features	auto, $\sqrt{\mathbf{N}}$, $\log N$
Profundidade máx.	$\sqrt{\mathbf{N}}$, N
Regressão Logística	
Hiper-Parâmetro	**Valores**
C	0.001, **0.01**, 0.1, 1, 10, 100, 1000
Penalidade	**l1**, l2
Gradient Boosting	
Hiper-Parâmetro	**Valores**
Taxa aprendizado n	0.001, 0.01, **0.1**, 1
Profundidade máx.	3, 4, **5**, 6
XGBoost	
Hiper-Parâmetro	**Valores**
Taxa aprendizado n	0.001, 0.01, **0.1**, 1
Profundidade máx.	3, 4, **5**, 6

Tabela 5.4: Hiper-parâmetros selecionados para os modelos

Método		Logistic Regression	Random Forest	Knn	GradBoost	XGBoost
F1	Média	0,3881	0,4431	0.4224	0,4568	0,4534
	D.P.	0,0122	0,0090	0,0097	0,0093	0,0071

Tabela 5.5: Resultados da aplicação de método

Para comparar o desempenho dos métodos, realizamos o teste-t de Student, para-par entre as execuções, cujos resultados para o *p-value* são apresentados na tabela 5.6. O teste-t é um outro método estatístico que avalia se existe diferença significativa entre as médias de duas populações de experimentos.

	LR	**RF**	**KNN**	**GradBoost**	**XGBoost**
LR	**x**	1×10^{-9}	8×10^{-7}	3×10^{-11}	2×10^{-11}
RF	1×10^{-9}	**x**	0,0003	0,0036	0,0105
KNN	8×10^{-7}	0,0003	**x**	4×10^{-7}	4×10^{-7}
GradBoost	3×10^{-11}	0,0036	4×10^{-7}	**x**	0,3779
XGBoost	2×10^{-11}	0,0105	4×10^{-7}	0,3779	**x**

Tabela 5.6: Valores p resultantes do teste t para cada par de experimentos

A análise da tabela 5.6 permite observar que as execuções, com exceção do par Gradient Boosting e XGBoost, são todas significativamente diferentes entre si com uma confiança de 95%, visto que os p-valores são menores que o limiar de 0,05. Dessa maneira, é possível escolher aquele valor com a maior média. Selecionamos o Gradient Boosting e o XGBoost como os melhores modelos dentre os aplicados, dado que ambos são estatisticamente iguais entre si e superiores em relação aos outros ao se comparar as médias.

As Tabelas 5.7 a 5.11 mostram as respectivas matrizes de confusão calculadas a partir das predições das publicações de usuários que apresentaram sequências de evolução da proficiência sem inconsistências, ou seja, que se mantiveram constantes ou crescentes. Também incluem a **revocação** (fração de instâncias relevantes que são recuperadas) e a **precisão** (fração de instâncias recuperadas que são relevantes), indicando que todos os modelos sofrem de um mesmo problema: a maioria das instâncias de cada classe tende a ser classificada como proficiência intermediária.

		Estimado			
		0	**1**	**2**	Revocação
Real	**0**	**555**	3154	265	14,0%
	1	302	**3119**	343	82,9%
	2	90	1069	**204**	19,3%
	Precisão	58,6%	42,5%	25,1%	

Tabela 5.7: Logistic Regression

		Estimado			
		0	**1**	**2**	Revocação
Real	**0**	**990**	2812	172	24,9%
	1	196	**3409**	159	90,6%
	2	88	948	**327**	24,0%
	Precisão	77,7%	47,6%	26,1%	

Tabela 5.8: Random Forest

		Estimado			
		0	**1**	**2**	Revocação
Real	**0**	**904**	2720	350	22,8%
	1	547	**2790**	427	74,1%
	2	188	931	**244**	17,9%
	Precisão	55,2%	43,3%	23,9%	

Tabela 5.9: KNN

		Estimado			
		0	**1**	**2**	Revocação
Real	**0**	**1845**	1885	244	46,4%
	1	570	**2903**	291	77,1%
	2	217	720	**426**	31,3%
	Precisão	70,1%	52,7%	44,3%	

Tabela 5.10: Gradient Boosting

		Estimado			
		0	**1**	**2**	Revocação
Real	**0**	**1681**	2041	252	42,30%
	1	577	**2889**	298	76,75%
	2	211	770	**382**	28,02%
	Precisão	68,08%	50,68%	43,51%	

Tabela 5.11: XGBoost

A análise das matrizes de confusão demonstra que, de fato, os modelos tendem a classificar a maioria das publicações como intermediário. Apesar disso, os modelos baseados em *boosting* apresentaram maior valor de revocação para as proficiências iniciante e avançada, corroborando com a conclusão de que esses são os melhores modelos dentre os escolhidos.

Dessa maneira, treinamos os modelos XGBoost e *Gradient Boosting*, que apresentaram valores de assertividade muito semelhantes, para estimar a probabilidade de cada texto apresentar cada um dos três níveis de proficiência e esses utilizamos esses resultados para executar o modelo sequencial de classificação.

5.6 SEMPLICe

A aplicação dos diferentes métodos supervisionados de classificação, demonstra que, a partir de características do texto, como quantidade de palavras em uma frase ou tamanho das palavras, é possível classificar o seu nível de proficiência, embora os valores de F1 obtidos não sejam muito altos. A comparação entre os resultados obtidos indica que os métodos que melhor descrevem a proficiência de um texto em alemão, baseado nas características passadas, são o Gradient Boosting e o XGBoost.

Portanto, utilizamos esses métodos para classificar a probabilidade de cada texto apresentar cada nível de proficiência (iniciante, intermediário ou avançado). Esses valores são insumo para o método sequencial de classificação, que considera também a ordem temporal das publicações: um usuário pode apenas manter ou aumentar seu

nível de proficiência, mas nunca diminuir.

Por isso, além de considerar a probabilidade de um post ser de determinada proficiência, esse modelo também observa se os posts anteriores do mesmo usuário também apresentavam a mesma ou menor proficiência. Os resultados, comparando o método alimentado pelo Gradient Boosting e XGBoost, estão apresentados a seguir.

5.6.1 Gradient Boosting

A Tabela 5.12 apresenta a comparação entre as aplicações do Gradient Boosting e do SEMPLICe utilizando os resultados deste modelo como insumo. Nela, apresentamos o F1 ponderado do Gradient Boosting (GB) e incluímos o respectivo resultado para o SEMPLICe quando as previsões do GB são usadas como insumo. Após o uso do algoritmo SEMPLICe, o F1 ponderado aumentou de 0,4568 para 0,5919, representando um ganho de 29,6% em relação ao melhor modelo não-sequencial em nossos experimentos. As publicações foram classificadas corretamente com uma precisão de 62,74%.

	Não-sequencial (GB)	GB+SEMPLICe
Média	0,4568	0,5919
Desvio	0,0093	0,0098

Tabela 5.12: Média e desvio padrão do F1 ponderado obtido pelo Gradient Boosting (GB) e pelo SEMPLICe.

A Tabela 5.13 contém a matriz de confusão para os resultados do SEMPLICe alimentado pelo GB. Em relação aos resultados do GB, observa-se que o SEMPLICe causou um aumento na revocação da classe intermediária de 77% para 94%, enquanto a classe iniciante teve uma pequena queda de 46% para 45%, e a classe avançada teve uma queda maior, descendo de 31% para 24%. No entanto, o número de erros grosseiros (classificar iniciante como avançado e vice-versa) reduziu de maneira drástica. Discutimos este resultado em detalhe na seção seguinte.

		Estimado			
		0	**1**	**2**	Revocação
Real	**0**	**1754**	2149	70	45,2%
	1	150	**3543**	70	94,2%
	2	63	976	**324**	23,8%
	Precisão	89,2%	53,1%	69,8%	

Tabela 5.13: Matriz de confusão do SEMPLICe usando Gradient Boosting como entrada.

5.6.2 XGBoost

Nesta seção, apresentamos a aplicação do SEMPLICe utilizando os resultados do XGBoost como insumo. A Tabela 5.14 apresenta a comparação entre essas aplicações. Nela, apresentamos o F1 ponderado do XGBoost (XGB) e do SEMPLICe. Após o uso do algoritmo SEMPLICe, o F1 ponderado aumentou de 0,4534 para 0,5385, o que representa um ganho de 18,8% em relação ao melhor modelo não-sequencial. As publicações foram classificadas corretamente com uma precisão de 57,62%.

	Não-sequencial (XGB)	XGB+SEMPLICe
Média	0,4534	0,5385
Desvio	0,0093	0,0098

Tabela 5.14: Média e desvio padrão do F1 ponderado obtido pelo XGBoost X(GB) e pelo SEMPLICe

A Tabela 5.15 apresenta a matriz de confusão para os resultados do SEMPLICe alimentado pelo XGB. É possível observar que o SEMPLICe também causou um aumento na revocação da classe intermediária de 77% para 94%, enquanto a classe iniciante teve uma queda de 42% para 37%, e a classe avançada de 28% para 18%. No entanto, o número de erros grosseiros também reduziu de maneira drástica.

		Estimado			
		0	**1**	**2**	Revocação
Real	**0**	**1464**	2441	68	36,92%
	1	148	**3539**	76	94,1%
	2	61	1062	**240**	17,6%
	Precisão	87,51%	50,26%	62,5%	

Tabela 5.15: Matriz de confusão do SEMPLICe usando XGBoost como entrada.

5.6.3 Análise da composição de erro do SEMPLICe

É importante ressaltar que o SEMPLICe quando recebe como entrada os resultados do GB apresentou valores melhores quando comparado ao modelo alimentado pelo XGB. Dessa maneira, realizamos uma nova análise no resultado dessa aplicação para melhor entender a composição do erro das previsões retornadas pela combinação GB+SEMPLICe.

Nessa análise, agrupamos os usuários conforme suas curvas de aprendizado, por exemplo, um usuário do tipo $1 \rightarrow 2$ é um usuário que em algum momento transicionou de iniciante para intermediário. A Tabela 5.16 mostra a taxa de acerto médio por

grupo. Observa-se que a maioria das publicações associadas à proficiência avançada (i.e., índice 3) correspondem a usuários que já entram na comunidade com tal nível. A taxa de acerto média para estas publicações é baixa (23,95%). Uma possível explicação para este fato é que a comunidade em estudo é voltada para o compartilhamento de informações entre usuários dos mais diversos níveis e, dessa maneira, usuários avançados podem buscar utilizar uma linguagem mais simples ao responder questões de usuários menos proficientes.

Outra observação interessante desses resultados é como é alta a assertividade para as sequências que seguem uma ordem natural (indo de um nível para outro gradualmente): $1 \rightarrow 2$, $1 \rightarrow 2 \rightarrow 3$ e $2 \rightarrow 3$. Essa observação reforça ainda mais o SEMPLICe como um bom modelo para classificar sequências de publicações no subreddit.

	Sequências de proficiências						
	1	2	3	$1 \rightarrow 2$	$1 \rightarrow 3$	$2 \rightarrow 3$	$1 \rightarrow 2 \rightarrow 3$
Total de usuários	606	417	132	43	3	16	2
Média de posts	11	11	11	10	14	11	9
Taxa de acerto médio	47,19%	89,19%	23,95%	66,73%	20,00%	61,23%	72,36%

Tabela 5.16: Estatísticas do SEMPLICe para usuários agrupados por sequência de proficiência (1: iniciante, 2: intermediário, 3: avançado, $\rightarrow$: há transição).

5.6.4 Conclusão

Os resultados obtidos pelo SEMPLICe demonstraram um ganho de 29,7% na F1 ponderada em relação às previsões originais do Gradient Boosting (GB), usadas como entrada. De maneira geral, o SEMPLICe obteve uma revocação bastante elevada para categoria intermediário, teve uma revocação maior ou igual a dos outros modelos para categoria iniciante e teve uma revocação menor que o GB para a categoria avançada, embora tenha classificado menos instâncias desta classe como iniciantes (erros grosseiros).

5.6.5 Limitações

Os dados utilizados podem conter imprecisões, porque não é garantido que o usuário escreveu o próprio texto (e.g., um usuário iniciante pode publicar um texto mais complexo, pedindo ajuda para tradução). Além disso, não é possível afirmar que os níveis de proficiência indicados pelos usuários estão corretos, dado que a proficiência é auto-declarada. Para minimizar este problema, pode-se estudar a possibilidade considerar como *ground-truth* apenas publicações cujas *tags* de proficiência são baseadas no Quadro Europeu Comum de Referência para Línguas (CEFR) (i.e., A1, A2, B1, B2,

C1 e C2) [Figueras, 2012], assumindo-se que elas se baseiam em resultados de testes oficiais de proficiência para indicar seu nível. É interessante, futuramente, realizar a classificação apenas com esses usuários e avaliar a assertividade dessa classificação.

Outra limitação dos dados é que são apenas textos curtos e informais, escritos na Web. Apesar da quantidade de textos, muitos podem conter gírias ou abreviações comuns à Web, o que pode prejudicar essa análise. Outros trabalhos nessa linha utilizaram textos de testes de proficiência, em que voluntários respondem questões com a finalidade de medir a proficiência. Textos de redes sociais, em contrapartida, não são escritos com essa finalidade, o que pode impactar diretamente nos resultados.

Apesar das limitações dos dados, que são inerentes às análises que utilizam dados de redes sociais online Pappa et al. [2017]; Cunha et al. [2016], o nosso modelo se mostrou muito bom para realizar a classificação de publicações, visto que elevou a acurácia dos classificadores para um bom valor de acurácia. Esse valor se equipara a outros trabalhos da literatura que também realizaram classificação da proficiência em idiomas, os quais não encontraram limitações como as do nosso conjunto de dados, visto que utilizaram textos previamente classificados por profissionais da área como entradas para treinar seus algoritmos Yang et al. [2016]; Crossley et al. [2012].

Capítulo 6

Conclusão

Esta dissertação teve como foco o estudo de comunidades do Reddit voltadas para aprendizado de idiomas estrangeiros. Inicialmente, analisamos o comportamento dos usuários, modelando as interações entre cada par de usuários como um grafo. Isso permitiu observar que os usuários em um subreddit se atentam mais ao conteúdo que às relações interpessoais. Esse fato foi corroborado pela análise das *threads*, que é o conjunto de respostas recebidas por cada publicação. Observamos que a maior parte das discussões são curtas, não se prolongando por muitas comentários.

Em seguida, analisamos o perfil dos usuários de uma dessas comunidades, voltada para o idioma alemão. A escolha dessa comunidade se deu por ser a única que pede aos usuários para classificarem sua proficiência ao participar do subreddit. Observamos que a maior parte dos usuários ativos são nativos e estão na comunidade para auxiliar usuários que ainda estão aprendendo o idioma. Apesar disso, ainda existem muitas interações de usuários que apresentam níveis de proficiência mais baixo.

No passo seguinte analisamos as publicações de usuários agrupados por nível de proficiência utilizando o LIWC, ferramenta para extrair características textuais das publicações. Nossos resultados indicaram que usuários iniciantes, intermediários e avançados apresentam características diferentes em seus textos.

Em seguida, utilizamos as características extraídas pelo LIWC como insumo para métodos de classificação supervisionada, de maneira a tentar inferir a proficiência de um usuário dado o texto publicado. Testamos cinco classificadores diferentes, tendo os métodos baseados em *boosting* (*Gradient Boosting* e XGBoost) apresentado os melhores resultados com relação à métrica F1 ponderada por classe. Apesar disso, os resultados ainda não tiveram um bom valor de assertividade.

Como principal contribuição deste trabalho, propusemos um método que permite melhorar as previsões feitas por um classificador de proficiência para publicações in-

dividuais ao agrupá-las por usuário e ordená-las no tempo. Este método assume que o nível de proficiência é uma função monotonicamente não-decrescente. Apresentamos um algoritmo força bruta para este método cuja complexidade é $\Theta(N^2)$ onde N é o número de publicações escritas por um usuário. Propusemos em seguida um algoritmo baseado no paradigma de programação dinâmica que retorna os mesmos resultados, mas com complexidade $\Theta(N)$. Este algoritmo foi denominado de SEMPLICe.

Através do SEMPLICe foi possível aprimorar as previsões retornadas pelos melhores classificadores da abordagem não-sequencial, elevando o F1 ponderado em 29.6%. O SEMPLICe também reduziu de forma substancial a ocorrência de erros grosseiros. A maioria dos erros cometidos pelo SEMPLICe são nas classificações de publicações escritas por usuários que já entram na comunidade como usuários avançados. Uma das hipóteses que explicaria este fenômeno e que pretendemos investigar futuramente é a de que tais usuários buscam utilizar uma linguagem mais simples ao interagir com usuários menos proficientes.

Apesar de ter alcançado um bom resultado, o SEMPLICe ainda pode ser aprimorado. Para isso, como trabalhos futuros, pretendemos realizar dois experimentos. O primeiro, consiste em realizar a retroalimentação do modelo: utilizar as sequências de proficiências retornadas pelo próprio SEMPLICe para alimentá-lo novamente, repetindo esse processo iterativamente até os resultados convirjam. O segundo método consiste em realizar a predição de quando ocorrem as transições de nível de proficiência utilizando uma rede neural recorrente, como o *Long Short Term Memory* (LSTM), que realizaria a propagação de informação (i.e., quando ocorrem as transições) ao longo da classificação das publicações.

Os resultados deste trabalho podem ser utilizados por usuários de mídias sociais voltadas para o aprendizado de idiomas, para acompanhar a evolução de sua proficiência ao longo do tempo. Por exemplo, as previsões do SEMPLICe podem ser utilizadas para auxiliar no direcionamento das atividades de uma comunidade de aprendizado de língua, baseado na proficiência de seus usuários. Trabalhos futuros incluem a realização de experimentos com outros idiomas, como espanhol e francês, assim como utilizar dados de outras redes além do Reddit. Adicionalmente, almeja-se a criação de um *bot* do *Reddit* que identifica quando um usuário provavelmente melhorou seu nível de proficiência e o notifica sobre isso.

Referências Bibliográficas

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175--185.

Andalibi, N.; Haimson, O. L.; De Choudhury, M. & Forte, A. (2016). Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. Em *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3906--3918. ACM.

Anderson, K. E. (2015). Ask me anything: what is reddit? *Library Hi Tech News*, 32(5):8--11.

Arnold, N. & Paulus, T. (2010). Using a social networking site for experiential learning: Appropriating, lurking, modeling and community building. *The Internet and higher education*, 13(4):188--196.

Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281--305.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.

Buntain, C. & Golbeck, J. (2014). Identifying social roles in reddit using network structure. Em *Proceedings of the 23rd International Conference on World Wide Web*, pp. 615--620. ACM.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Em *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785--794. ACM.

Chester, A. & O'Hara, A. (2007). Image, identity and pseudonymity in online discussions. *International Journal of Learning*, 13(12).

Choi, D.; Han, J.; Chung, T.; Ahn, Y.-Y.; Chun, B.-G. & Kwon, T. T. (2015). Characterizing conversation patterns in reddit: From the perspectives of content properties

and user participation behaviors. Em *Proceedings of the 2015 ACM on Conference on Online Social Networks*, pp. 233--243. ACM.

Clark, C. & Gruba, P. (2010). The use of social networking sites for foreign language learning: An autoethnographic study of livemocha. *Curriculum, technology & transformation for an unknown future. Proceedings ascilite Sydney*, pp. 164--173.

Crossley, S. A.; Salsbury, T. & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2):243--263.

Crossley, S. A.; Salsbury, T.; McNamara, D. S. & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28(4):561--580.

Cunha, T. O.; Weber, I.; Haddadi, H. & Pappa, G. L. (2016). The effect of social feedback in a reddit weight loss community. Em *Proceedings of the 6th International Conference on Digital Health Conference*, pp. 99--103. ACM.

De Choudhury, M. & De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. Em *ICWSM*.

Duggan, M. & Smith, A. (2013). 6% of online adults are reddit users. *Pew Internet & American Life Project*, 3:1--10.

Figueras, N. (2012). The impact of the cefr. *ELT journal*, 66(4):477--485.

Flanagan, B.; Hirokawa, S.; Kaneko, E. & Izumi, E. (2016). Classification of speaking proficiency level by machine learning and feature selection. Em *International Symposium on Emerging Technologies for Education*, pp. 677--682. Springer.

Fraga, B. S.; da Silva, A. P. C. & Murai, F. (2018). Online social networks in health care: A study of mental disorders on reddit. Em *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 568--573. IEEE.

Freedman, D. A. (2009). *Statistical models: theory and practice*. cambridge university press.

Friedman, J.; Hastie, T. & Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189--1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367--378.

Gagnon, T. (2013). The disinhibition of reddit users. *Adele Richardson's Spring*.

Hand, D. & Christen, P. (2018). A note on using the f-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3):539--547.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Em *Ijcai*, volume 14, pp. 1137--1145. Montreal, Canada.

Kou, Y.; Gray, C.; L Toombs, A. & S Adams, R. (2018). Knowledge production and social roles in an online community of emerging occupation: A study of user experience practitioners on reddit.

Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Lin, C.-H.; Warschauer, M. & Blake, R. (2016). Language learning through social networks: Perceptions and reality.

Mills, R. A. (2015). Reddit. com: A census of subreddits. Em *Proceedings of the ACM Web Science Conference*, p. 49. ACM.

Newman, M. E. (2011). Complex systems: A survey. *arXiv preprint arXiv:1112.1440*.

Olson, R. S. & Neal, Z. P. (2015). Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science*, 1:e4.

Pappa, G. L.; Cunha, T. O.; Bicalho, P. V.; Ribeiro, A.; Silva, A. P. C.; Meira Jr, W. & Beleigoli, A. M. R. (2017). Factors associated with weight change in online weight management communities: A case study in the loseit reddit community. *Journal of medical Internet research*, 19(1).

Saramäki, J.; Kivelä, M.; Onnela, J.-P.; Kaski, K. & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105.

Schmidt, M.; Le Roux, N. & Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83--112.

Silveira, B.; da Silva, A. P. C. & Murai, F. (2018). Análise de comunidades de suporte a transtornos de saude mental do reddit. Em *7º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)*, volume 7. SBC.

Singer, P.; Flöck, F.; Meinhart, C.; Zeitfogel, E. & Strohmaier, M. (2014). Evolution of reddit: from the front page of the internet to a self-referential community? Em *Proceedings of the 23rd International Conference on World Wide Web*, pp. 517--522. ACM.

Spolaôr, N. & Tsoumakas, G. (2013). Evaluating feature selection methods for multi-label text classification. *BioASQ workhsop*.

Suler, J. R. (2002). Identity management in cyberspace. *Journal of applied psychoanalytic studies*, 4(4):455--459.

Warschauer, M. & Healey, D. (1998). Computers and language learning: An overview. *Language teaching*, 31(2):57--71.

Weninger, T.; Zhu, X. A. & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. Em *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 579--583. ACM.

Yang, Y.; Yu, W. & Lim, H. (2016). Predicting second language proficiency level using linguistic cognitive task and machine learning techniques. *Wireless Personal Communications*, 86(1):271--285.

Yu, H.-F.; Huang, F.-L. & Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41--75.

Zechner, K. & Bejar, I. I. (2006). Towards automatic scoring of non-native spontaneous speech. Em *Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics*, pp. 216--223. Association for Computational Linguistics.

Zhao, Y. (1996). Language learning on the world wide web: Toward a framework of network based call. *Calico Journal*, pp. 37--51.

Zourou, K. (2012). De l'attrait des médias sociaux pour l'apprentissage des langues–regard sur l'état de l'art. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 15(1).