

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Avaliação da qualidade da Wikipédia enquanto fonte de informação em saúde

Luís Couto



Mestrado Integrado em Engenharia Informática e Computação

Orientador: Carla Teixeira Lopes

Co-orientador: Gil Domingues

Julho de 2021

Avaliação da qualidade da Wikipédia enquanto fonte de informação em saúde

Luís Couto

Mestrado Integrado em Engenharia Informática e Computação

Julho de 2021

Abstract

Wikipedia is an online, free, multi-idiom, and collaborative encyclopedia. Nowadays, it is one of the largest sources of online knowledge, often appearing at the top of the results of the major search engines. There, it is possible to find information from different areas, from technology to philosophy, including health.

As a health-related data source, it is one of the most used sources of information, used not only by the general public but also by professionals. The reason for such a broad public is that, apart from the content of the articles, it includes external links for additional data sources as well. Despite being a top-rated resource, the open nature of Wikipedia contributions, where there are no curators, raises safety concerns, specifically in the health context, as such data is used for decision-making. There are, however, many discrepancies among the Wikipedia versions for all available idioms. These differences can be an obstacle to people's equal access to information. Thus, it is crucial to evaluate the information and compare the various idioms in this regard.

In the first stage, the quality of health-related Wikipedia articles across different languages was compared. Specifically, in articles available in languages with over one hundred million speakers, and also in Catalan, Greek, Italian, Korean, Turkish, Perse, and Hebrew, for its historical tradition. A set of predefined metrics was used to evaluate the authority, completeness, complexity, informativeness, consistency, currency, and volatility of health information on Wikipedia. These quality metrics were applied to a list of articles in English, maintained by WikiProject Medicine, and their versions in the above idioms. After that, in a second stage, general and health-specific features from Wikipedia articles were used to propose health-specific metrics. Finally, these metrics were posteriorly evaluated using a set of English articles previously assessed by WikiProject Medicine.

Differences in the quality of information between idioms were demonstrated. English is the idiom with the highest quality in general. Urdu, Greek, Indonesian, and Hindi achieved lower values of quality. A correlation was also identified between the quality of information and the number of speakers and number of articles available on Wikipedia. With that, this dissertation contributes to a clarification of Wikipedia's discrepancies in access to health information. We combined generic and specific features to propose four specific metrics – HealthAuthority, HealthCompleteness, HealthInformativeness, and HealthConsistency. We concluded that, by combining generic and specific features, we can improve the assessment of the quality of health-related information on Wikipedia. Along with the explored features, these metrics can be used in approaches that automatically classify the quality of Wikipedia health-related articles.

Keywords: Information quality, Wikipedia, Health information, Multi-idiom information access.

Resumo

A Wikipédia é uma enciclopédia *online*, gratuita, multi-idioma e colaborativa. Atualmente, é uma das maiores fontes de conhecimento *online*, aparecendo muitas vezes no topo dos resultados dos principais motores de busca. É possível encontrar, na Wikipédia, informações de diversas áreas, desde tecnologia à filosofia, passando pela saúde.

Como fonte de informação de saúde, é uma das mais utilizadas, não só pelo público em geral, mas também pelos profissionais. A razão para um público tão amplo é que, além do conteúdo dos artigos, inclui *links* externos para fontes de informação adicionais. Apesar de ser um recurso de primeira linha, a natureza aberta das contribuições da Wikipédia, onde não há curadores, levanta questões de segurança, especificamente no contexto da saúde, uma vez que a informação é usada para a tomada de decisões. Existem, no entanto, muitas discrepâncias entre as versões da Wikipédia para todos os idiomas disponíveis. Essas diferenças podem ser um obstáculo ao acesso equitativo das pessoas à informação. Assim, é fundamental avaliar a qualidade da informação e comparar os diversos idiomas a esse respeito.

Na primeira fase, foi comparada a qualidade de artigos da Wikipédia relacionados com saúde em diferentes idiomas. Mais especificamente, em artigos disponíveis em idiomas com mais de cem milhões de falantes, e também em catalão, grego, italiano, coreano, turco, persa e hebreu, pela sua tradição histórica. Foi utilizado um conjunto de métricas predefinidas para avaliar a *authority*, *completeness*, *complexity*, *informativeness*, *consistency*, *currency*, e *volatility* da informação na Wikipédia. Essas métricas de qualidade foram aplicadas a uma lista de artigos em inglês, mantida pelo WikiProject Medicine, e nos idiomas acima referidos. Posteriormente, numa segunda fase, foram exploradas medidas genéricas e específicas dos artigos de saúde da Wikipédia, para propor métricas específicas de saúde. Posteriormente, essas métricas foram avaliadas, utilizando um conjunto de artigos em inglês previamente avaliados pelo WikiProject Medicine.

Foram demonstradas diferenças na qualidade da informação entre os idiomas. O inglês é o idioma com maior qualidade em geral. O urdu, o grego, o indonésio e o hindi alcançaram os valores de qualidade mais baixos. Também foi identificada uma correlação entre a qualidade da informação e o número de falantes e de artigos disponíveis na Wikipedia. Com isto, esta dissertação contribui para o esclarecimento das discrepâncias da Wikipédia no acesso à informação em saúde. Combinámos medidas genéricas e específicas para propor quatro métricas específicas - HealthAuthority, HealthCompleteness, HealthInformativeness e HealthConsistency. Concluimos que, combinando medidas genéricas e específicas, podemos melhorar a avaliação da qualidade da informação relacionada com saúde na Wikipédia. Em conjunto com as medidas exploradas, essas métricas podem ser usadas em abordagens que classificam automaticamente a qualidade dos artigos da Wikipédia relacionados com saúde.

Palavras-chave: Qualidade da informação, Wikipédia, Informação em saúde, Pesquisa multi-idioma.

Agradecimentos

Primeiro, quero agradecer à minha orientadora, Professora Doutora Carla Teixeira Lopes, pela forma como me orientou, com constante disponibilidade, e pela motivação transmitida. Queria ainda agradecer ao meu co-orientador, Gil Domingues, pelo apoio, sobretudo no arranque do trabalho.

Em seguida, quero agradecer à FEUP por todo o percurso que me proporcionou, e em particular ao DEI, pela oportunidade de participar na The Web Conference 2021.

Em seguida, os meus agradecimentos vão para a minha família, pelo apoio ao longo de todo o percurso, com um pedido de desculpas pelo tempo que lhes roubei.

Quero ainda agradecer ao meu amigo Tiago, pelo incentivo que me ajudou a iniciar este percurso e a terminá-lo.

Foi uma jornada longa, talvez tardia, mas muito gratificante.

Obrigado a todos.

Luís Couto

*“All knowledge is connected to all other knowledge.
The fun is in making the connections.”*

Arthur C. Aufderheide

Conteúdo

1	Introdução	1
1.1	Contexto	1
1.2	Motivação e objetivos	2
1.3	Contribuições	2
1.4	Estrutura do documento	3
2	A Wikipédia	5
2.1	Breve história	5
2.2	Recomendações para a elaboração de conteúdos	6
2.3	Fiabilidade da Wikipédia	7
2.4	Mecanismos internos de qualidade	9
2.5	Diferenças entre idiomas	13
2.5.1	Artigos e utilizadores da Wikipédia	15
2.5.2	Qualidade do conteúdo	16
2.6	Informação de saúde na Wikipédia	17
2.6.1	WikiProject Medicine	18
3	Qualidade da informação existente na Wikipédia	21
3.1	Introdução	21
3.2	O conceito de qualidade da informação	22
3.3	Qualidade em geral	22
3.4	Qualidade em saúde	25
3.5	Análise comparativa	26
4	Qualidade dos conteúdos multi-idioma de saúde na Wikipédia	31
4.1	Formulação do problema	31
4.1.1	Comparação da qualidade da Wikipédia em diferentes idiomas	31
4.1.2	Avaliação automática da qualidade	32
4.2	Solução	32
4.2.1	Métricas de avaliação da qualidade	32
4.3	Metodologia	35
4.4	Elaboração do <i>dataset</i>	36
4.4.1	Seleção dos idiomas	36
4.4.2	Recolha dos artigos multi-idioma	38
4.4.3	Recolha de dados para as métricas específicas	38
4.4.4	Caraterização do <i>dataset</i>	39

5	Comparação multi-idioma da qualidade de artigos de saúde da Wikipédia	41
5.1	Cálculo de medidas e métricas	41
5.2	Análise estatística	43
5.2.1	Estratégia de análise	43
5.2.2	Resultados	43
5.2.3	Discussão de resultados	53
5.3	Conclusão	54
6	Proposta de medidas e métricas de avaliação específicas de saúde	57
6.1	Proposta de medidas específicas de saúde da Wikipédia	57
6.1.1	Número de <i>templates</i> de saúde	57
6.1.2	Número de valores nas <i>infoboxes</i> de saúde	58
6.1.3	Número de imagens nas <i>infoboxes</i> de saúde	59
6.1.4	Rácio de edições de administradores do Wikiproject Medicine	59
6.1.5	Tradução pela Healthcare Translation Task Force	59
6.1.6	Número de códigos médicos	59
6.1.7	Número de <i>links</i> reputados	59
6.1.8	Número de secções recomendadas	60
6.2	Análise de medidas de avaliação da qualidade	60
6.2.1	Análise de medidas genéricas	60
6.2.2	Análise de medidas específicas de saúde	61
6.3	Proposta de métricas específicas	63
6.3.1	Importância das medidas nas métricas genéricas	63
6.3.2	Proposta de métricas específicas de saúde	64
6.4	Avaliação das métricas de avaliação	65
6.5	Conclusão	66
7	Conclusões e trabalho futuro	67
7.1	Conclusões	67
7.2	Trabalho futuro	68
	Referências	69

Lista de Figuras

2.1	Hierarquia de grupos na Wikimedia	11
3.1	Estimativa do tamanho de uma versão impressa da Wikipédia	21
3.2	Mapeamento entre três modelos de avaliação	24
4.1	Metodologia utilizada	35
4.2	Estrutura da base de dados	39
4.3	Número de artigos por idioma	40
5.1	Distribuição das medidas da métrica <i>authority</i>	45
5.2	Distribuição das medidas da métrica <i>completeness</i>	46
5.3	Distribuição da medida Flesch reading ease	47
5.4	Distribuição das medidas da métrica <i>informativeness</i>	48
5.5	Distribuição das medidas da métrica <i>consistency</i>	50
5.6	Distribuição da medida do tempo médio de reversão das edições	52
5.7	Distribuição da medida atualidade	53
6.1	<i>Template</i> do artigo sobre COVID-19 na Wikipédia lusófona	58
6.2	<i>Infobox</i> do artigo sobre COVID-19 na Wikipédia lusófona	58
6.3	Distribuição das medidas específicas de saúde por nível de qualidade	62

Lista de Tabelas

2.1	Esquema de avaliação de artigos da Wikipédia	14
2.2	Estatísticas para versões da Wikipédia com mais de 1 milhão de artigos	15
3.1	Medidas utilizadas nos trabalhos analisados	28
4.1	Número de falantes de cada idioma do <i>dataset</i>	37
5.1	Processo de cálculo das medidas	42
5.2	Avaliação da qualidade dos idiomas para a métrica <i>authority</i>	44
5.3	Avaliação da qualidade dos idiomas para a métrica <i>completeness</i>	46
5.4	Avaliação da qualidade dos idiomas para a métrica <i>complexity</i>	47
5.5	Avaliação da qualidade dos idiomas para a métrica <i>informativeness</i>	48
5.6	Análise da <i>informativeness</i> sem aplicação de <i>stemming</i>	49
5.7	Avaliação da qualidade dos idiomas para a métrica <i>consistency</i>	50
5.8	Avaliação da qualidade dos idiomas para a métrica <i>volatility</i>	51
5.9	Avaliação da qualidade dos idiomas para a métrica <i>currency</i>	52
5.10	<i>Ranking</i> dos idiomas	54
5.11	Correlação entre as métricas e o número de falantes e de artigos	54
6.1	Medianas de medidas genéricas e sua correlação com a qualidade	61
6.2	Medianas das medidas específicas e sua correlação com a qualidade	62
6.3	Correlação das métricas com as suas medidas	63
6.4	Correlação entre métricas genéricas e específicas	65

Acrónimos

WMF	Wikimedia Foundation
IP	Internet Protocol
UNESCO	United Nations Educational, Scientific and Cultural Organization
FUNREDES	Fundación-Redes-y-Desarrollo
API	Application Programming Interface
EUA	Estados Unidos da América
WHO	World Health Organization
IIQ	Intervalo intercuartil
ISI	Idiomas significativamente inferiores
HTF	Healthcare Translation Task Force

Capítulo 1

Introdução

Neste primeiro capítulo, será apresentado, primeiramente, o contexto desta dissertação. São, posteriormente, identificados a motivação e os objetivos da mesma. São, ainda, expostas as contribuições, e o capítulo termina com uma descrição da estrutura deste documento.

1.1 Contexto

A Wikipédia é uma enciclopédia online, não comercial, de conteúdo aberto e gratuito, baseada num modelo colaborativo, significando que é acessível para consulta, mas também para edição imediata. Está actualmente disponível em 321 idiomas, com 310 deles activos [77] e conta actualmente com um total de mais de 400 milhões de artigos [72]. Isto faz com que seja dos *sites* mais visitados na *web*, somando atualmente 519 mil milhões de visualizações dos seus artigos [1]. Os motores de pesquisa *online* apresentam frequentemente no topo dos seus resultados artigos da Wikipédia [33], contribuindo, assim, para o elevado número de visualizações. Enquanto maior fonte de conhecimento online, a Wikipédia é utilizada por tecnologias e aplicações, como seja o caso dos assistentes virtuais.

Um estudo realizado pela Health On the Net Foundation [49] mostra que quando as necessidades de informação se relacionam com saúde, 44% dos inquiridos admitiu procurar por esta informação mais de três vezes por semana, sendo o principal ponto de acesso os motores de busca, o que poderá acabar por levá-los posteriormente à Wikipédia [31]. De referir, ainda, que três dos dez termos mais procurados no motor de busca Google no ano de 2020 [22], são relacionados com a saúde – “coronavirus”, “coronavirus update” e “coronavirus symptoms”. A Wikipédia conta já com mais de 40 mil artigos relacionados com saúde [26] apenas na sua versão inglesa, sendo que o conjunto dos artigos da área médica mais visitados na Wikipédia conta atualmente com mais de dois mil milhões de visitas anuais e o artigo mais popular, atualmente “COVID-19 pandemic” conta com uma média superior a 40 mil visitas diárias [91]. Esta importância é ainda acentuada pelo facto da Wikipédia ser também utilizada como fonte de informação, quer por estudantes de medicina [50, 45], quer por profissionais de saúde [66, 29, 53], exercendo, portanto, influência na sua prática clínica, e por conseguinte, na saúde dos utentes. O estudo da Health On the Net

Foundation [49] revelou, ainda, que a qualidade da informação continua a ser a maior barreira encontrada pelos inquiridos (80%) aquando da procura de informação de saúde online e os factores mais valorizados pelos mesmos relativamente à qualidade da informação prendem-se com a confiabilidade/credibilidade (96%) e com a precisão e disponibilidade (95%) da mesma. Dada a importância da qualidade da informação online e sua avaliação, em particular na Wikipédia, têm sido vários os estudos elaborados sobre este assunto desde a sua criação. De uma forma geral estes podem dividir-se, de acordo com os seus métodos, em avaliação manual e avaliação automática da qualidade. Dos estudos envolvendo diferentes métricas de avaliação da qualidade de artigos, uma parte deles incide em particular sobre a área da saúde, sendo estes estudos aqui analisados posteriormente. Os resultados da avaliação da qualidade variam, com alguns mais favoráveis e outros mais desfavoráveis, gerando uma crescente preocupação com a melhoria da qualidade da informação na área da saúde [55].

1.2 Motivação e objetivos

Dada a importância que a Wikipédia pode assumir enquanto fonte de informação relacionada com saúde, assume particular relevo a disponibilidade e qualidade desta informação a falantes dos diversos idiomas. Torna-se importante confirmar a assimetria sugerida pelas diferenças na quantidade de artigos disponíveis [77], que provavelmente reflete a mesma situação na restante *web*. Duas das metas definidas pela Division for Sustainable Development Goals da Organização das Nações Unidas, descritas na sua “Transforming our world: the 2030 Agenda for Sustainable Development”, são: “garantir uma vida saudável e promover o bem-estar para todos em todas as idades” (meta 3) e “garantir uma educação de qualidade inclusiva e equitativa e promover oportunidades de aprendizagem ao longo da vida para todos” (meta 4) [8].

Esta dissertação pode contribuir para as metas referidas, identificando possíveis assimetrias no acesso à informação. Com a identificação e definição do problema, dá-se o primeiro passo para a democratização do acesso à informação, em particular na área da saúde. Pretende-se, por outro lado, identificar características dos artigos de saúde da Wikipédia que possam ser utilizadas para melhorar as métricas já existentes de avaliação automática da qualidade da informação.

1.3 Contribuições

Com o esclarecimento das assimetrias existentes, esta dissertação, serve como incentivo à redução das diferenças no acesso à informação *online*, nomeadamente no que toca à saúde. Outra das contribuições são as medidas e consequentes métricas específicas, propostas, para avaliação da qualidade da informação em saúde na Wikipédia. Com o contributo desta proposta de medidas e métricas específicas, foi publicado um artigo científico – Assessing the quality of health-related Wikipedia articles with generic and specific metrics, na The Web Conference 2021, enquadrado no Wiki Workshop 2021. Foi ainda submetido outro artigo científico no âmbito do “17th International Symposium on Open Collaboration (OpenSym 2021)” – Equal opportunities in the access to

quality online health information? A multi-lingual study on Wikipedia, a aguardar aprovação. Este segundo artigo é baseado no contributo do esclarecimento das assimetrias no acesso à informação nos diferentes idiomas da Wikipédia. São ainda disponibilizados os *datasets* num repositório de dados de investigação institucional, relativos à primeira¹ e segunda² fases do trabalho. No mesmo repositório está ainda o código desenvolvido nesta dissertação.

1.4 Estrutura do documento

O presente documento encontra-se dividido em sete capítulos. Depois da introdução, no Capítulo 1, o Capítulo 2, faz um enquadramento geral da Wikipédia, desde a sua história, passando pelo conjunto de políticas que a caracterizam, uma análise à sua fiabilidade, às diferenças entre os diferentes idiomas em que está disponível, terminando na análise da área da saúde dentro da Wikipédia. No Capítulo 3 é feita uma revisão de trabalhos anteriores, com foco especial nas métricas utilizadas, quer para avaliação da qualidade em geral, quer para avaliação da área da saúde em específico, dentro da Wikipédia. O Capítulo 4 relata o problema formulado, a solução para o mesmo, a metodologia para lá chegar e descreve o *dataset* resultante. O Capítulo 5 descreve uma primeira primeira fase, de avaliação da qualidade de artigos de saúde da Wikipédia. No Capítulo 6 são propostas métricas e respetivas medidas específicas para avaliação da qualidade de artigos de saúde. Por fim, no Capítulo 7, são tecidas conclusões da dissertação e propostas sugestões de trabalho futuro.

¹<https://doi.org/10.25747/ep0v-en19>

²<https://doi.org/10.25747/wfzk-h937>

Capítulo 2

A Wikipédia

Neste capítulo, depois de uma descrição da história da Wikipédia, serão analisadas as suas principais políticas com influência na qualidade. Em seguida, é abordada a questão da fiabilidade da informação disponibilizada, abordando os mecanismos internos para assegurar níveis de qualidade consistentes. Posteriormente, serão analisadas as diferenças evidenciadas entre as versões de diferentes idiomas, quer do ponto de vista quantitativo, que do ponto de vista qualitativo. Por fim, será destacada a área da saúde dentro da Wikipédia, fazendo uma breve apresentação do WikiProject Medicine, motor de desenvolvimento da área médica na Wikipédia.

2.1 Breve história

Em Março de 2000, o empresário Jimmy Donal Wales [11] e o doutorando em filosofia Larry Sanger [30] criaram uma enciclopédia *online* livre – a Nupedia, escrita por voluntários especializados nas áreas sobre os conteúdos que editavam. Estas edições passavam posteriormente por um processo de revisão e aprovação, composto por sete passos que, pelo seu rigor e lentidão, se mostrou prejudicial para o projeto, que acabou por terminar em 2003, com apenas 23 artigos finalizados. Derivado deste projeto e apenas 10 meses após o início do mesmo, surgiu outro – a Wikipédia, lançada em 15 de Janeiro de 2001, em inglês.

Recorrendo ao conceito de *wiki*, desenvolvido em 1994 por Howard Cunningham, e assente numa criação de conteúdos baseado em *markup language*, a Wikipédia aproveita a possibilidade dos utilizadores poderem editar com relativa facilidade as páginas que desejam. São utilizadas páginas organizadas simples, com *hyperlinks*, diretamente na *web*, de forma descentralizada, permitindo uma edição de conteúdo simples e imediata. Estas características que diferenciavam claramente a Wikipédia da Nupedia, permitiram-lhe uma evolução e um sucesso completamente diferentes e, assim, no final do primeiro mês já contava com cerca de 600 artigos e, no final do primeiro ano, superou os 20 mil artigos. O segundo idioma, após o inglês, foi o alemão, seguido imediatamente pelo catalão, mantendo-se estes como os únicos idiomas durante dois meses. Houve, posteriormente, a inclusão de outros idiomas e, no final do ano, a Wikipédia tinha artigos escritos em 18 idiomas diferentes. Esta diversidade de idiomas, levou a que, de 90% de artigos escritos em

inglês no início de 2002, este valor tenha reduzido para menos de 50% do total volvidos apenas dois anos e, atualmente, se situe nos 11,2% [84].

2.2 Recomendações para a elaboração de conteúdos

Apesar do grande número de artigos criados inicialmente pelos primeiro utilizadores, estes não apresentavam a qualidade desejada, o que levou Larry Sanger a definir regras que publicou na própria Wikipédia nas páginas “Wikipedia is not a dictionary” [88] e “What wikipedia is not” [86], que ainda hoje existem, com mudanças sofridas ao longo do tempo. As políticas e as recomendações da Wikipédia são concebidas pelos seus utilizadores enquanto padrões que devem ser seguidos, sem no entanto se quererem impor como regras fixas. Neste contexto, surgem com particular relevo cinco princípios, considerados pilares, para definir de forma geral as regras e recomendações para a elaboração dos conteúdos [81].

O primeiro pilar, define que “a Wikipédia é uma enciclopédia”. Este pilar pretende definir o escopo da Wikipédia, sendo a sua definição feita sobretudo pela negativa, por aquilo que não se pretende que a Wikipédia seja, de modo a encaminhar a sua essência para o de uma enciclopédia generalista, especializada, ou almanaque.

O segundo pilar, indica que “a Wikipédia rege-se pela imparcialidade”. Este pilar manifesta-se particularmente importante, oriundo já dos tempos da Nupedia, onde se intitulava de “Non-bias policy”, e indica que os artigos da Wikipédia devem refletir as diferentes visões sobre o problema, incluindo-as e identificando claramente onde as disputas ocorrem, identificando os seus lados.

O terceiro pilar, estabelece que “a Wikipédia é uma enciclopédia de conteúdo livre”. Este pilar está ligado, sobretudo, a direitos de autor, e indica que não só a edição está acessível a qualquer um, como os textos estão também disponíveis a todos, licenciados sob os termos das licenças “Creative Commons – Atribuição – Compartilhamento pela mesma licença 3.0 Não Adaptada”¹ e da “Licença GNU de Documentação Livre”².

O quarto pilar, define que “a Wikipédia possui normas de conduta”. Este pilar refere-se, sobretudo, a princípios de etiqueta, a respeitar entre os vários utilizadores, evitando conflitos e mantendo um ambiente cordial.

O último pilar, estipula que “a Wikipédia não possui regras fixas além dos cinco pilares”. Partindo do pressuposto que “a exceção confirma a regra”, este pilar procura evitar uma rigidez nas regras que acabe por prejudicar o desenvolvimento da Wikipédia, apelando em vez disso ao senso comum, ao espírito das regras, ao consenso, clareza e flexibilidade, desde que haja uma justificação válida.

Estes cinco pilares são comuns às Wikipédias dos diferentes idiomas, existindo, no entanto, políticas definidas para cada versão, criadas pela comunidade, por consenso ou por votação, assumindo um carácter transversal a todos os artigos presentes e a todos os seus utilizadores, com

¹<https://creativecommons.org/licenses/by-sa/3.0/>

²<http://www.gnu.org/licenses/fdl-1.3.html>

possíveis sanções para os transgressores das mesmas, como seja o bloqueio do utilizador na plataforma por determinado período de tempo [97].

2.3 Fiabilidade da Wikipédia

Um estudo realizado em 2010 [54] analisou o comportamento de um tipo específico de vândalos da Wikipédia – os *trolls*: “indivíduos que colocam mensagens ou comentários provocadores, maldosos ou violentos em sítios de discussão pública on-line, com intuito desestabilizador”³, sendo que o vandalismo é proibido na Wikipédia. Usando uma das fraquezas da Wikipédia – o facto de qualquer um poder criar, editar ou apagar conteúdo, estes indivíduos vandalizam os conteúdos por motivações, segundo o estudo, de tédio, busca de atenção, vingança e diversão. As suas ações passam sobretudo por violações das políticas já aqui referidas, muitas vezes com intuito de destruição do espírito comunitário que caracteriza a Wikipédia, atuando normalmente sozinhos e de forma anónima. Existem no entanto outros tipos de vândalos e vandalismo a aproveitarem-se das fraquezas da Wikipédia, com diversas motivações mas quase sempre o mesmo efeito prático, no que à qualidade diz respeito.

O vandalismo é uma das maiores preocupações na manutenção da Wikipédia, dando inclusivamente origem a projetos anti-vandalismo⁴, onde voluntários se propõem a lutar contra este problema, com diversas ações, incluindo listas de artigos sob vigilância ativa e ferramentas criadas para auxiliar nesta luta. Tudo isto permite que os atos de vandalismo possam ser debelados, na maioria das vezes em 2 a 3 minutos [14].

O vandalismo não é o único motivo que levanta preocupação sobre a qualidade da informação, que é, de resto, assumida e declarada numa página da própria Wikipédia [83]. É referido que existe sempre a possibilidade de ser adicionada informação errada de forma intencional ou não, uma vez que não há qualquer triagem sobre os editores e edições da informação e os seus conhecimentos sobre o assunto a editar. Uma vez que não existe garantia dessa informação ser revista por alguém com mais conhecimentos, não existe assim qualquer garantia sobre a qualidade da mesma. Em 2010, a política francesa Ségolène Royal, no Dia Nacional de Recordação do Tráfico, da Escravidão e da Abolição, em França, prestou tributo a uma personagem histórica, cuja biografia tinha dado origem a um artigo da Wikipédia francesa já com 3 anos. Tratava-se de um cientista, humanista e naturalista, que sendo filho de um traficante de escravos, se recusou a seguir o caminho do seu pai, reprovando o tráfico de escravos. Tendo este tributo ecoado internacionalmente⁵, a questão levantada é que o personagem nunca existiu na realidade e faz parte atualmente da lista de artigos falsos da Wikipédia francesa, uma lista que cada Wikipedia mantém para evitar que se criem novamente artigos sobre estas temáticas ficcionais⁶. Esta história, apenas uma no meio de tantas, como demonstram as referidas listas de artigos falsos, por um lado expõe uma

³“trol”, in Dicionário Priberam da Língua Portuguesa, <https://dicionario.priberam.org/trol>

⁴<https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Projetos/AntiVandalismo>

⁵<https://www.sudouest.fr/2010/06/07/leon-robert-de-l-astran-celui-qui-n-a-jamais-existe-110539-7.php>

⁶https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Liste_de_canulars

das maiores fragilidades da Wikipédia, mas por outro, demonstra também a sua importância e dos seus conteúdos enquanto fonte de informação.

Outro aspeto acerca dos artigos presentes na Wikipédia é o facto destes poderem ser incompletos, *i.e.*, sem qualidade enciclopédica e que na Wikipédia podem ser definidos como “esboço”, ou então a informação presente poder simplesmente ser apagada por qualquer um, mais uma vez, de forma intencional ou não. Inversamente, também pode ser adicionada grande quantidade de texto nos artigos, tornando-os praticamente ilegíveis. A quantidade excessiva de informação também pode estar ligada à quantidade excessiva de artigos sobre temáticas em particulares, de interesse apenas para pequenos nichos de utilizadores e que, para os restantes será visto como inútil, tendo sido criado inclusivamente um termo para designar este acontecimento – *Fancruft*⁷.

A Wikipédia, baseada em texto, corre assim os riscos associados à qualidade da escrita em si, da linguagem utilizada, o que pode levantar problemas de qualidade gramatical ou ortográfica, de acordo com o utilizador que edita o texto; corre também os riscos associados à utilização de uma linguagem demasiado técnica, como o uso de jargões, que a tornam incompreensível à generalidade dos utilizadores. Ciente destas questões, a Wikipédia sugere ao utilizador, maneiras de evitar ou minimizar estes problemas, como seja a indicação de utilizar múltiplas fontes independentes, de fazer uma avaliação das fontes citadas nos artigos e, em última instância, utilizar os artigos como ponto de partida na busca de informação, ou como uma vista geral da mesma. É, ainda, aconselhada uma avaliação ao histórico das edições dos artigos e discussões sobre estes, para melhor entender o porquê da presença de alguma da informação. Fazendo uso de uma das características da Wikipédia, as hiperligações, estas são aconselhadas como forma de obter informação complementar dentro da própria Wikipédia. Adicionalmente é, ainda, referida a vantagem da divisão da informação em categorias e da ajuda que a comunidade ativa pode dar a todos os utilizadores.

As causas descritas para os problemas são, por outro lado, a causa para algumas das vantagens da Wikipédia [87], nomeadamente no que diz respeito à qualidade da informação. O facto de poder ser rápida e facilmente editada, que pode ser utilizada para degradar a qualidade, é também utilizado para reparar os danos de forma rápida e fácil, significando que umas das suas características estruturais é o facto de estar constantemente atualizada. Enquanto que, edições tradicionais de enciclopédias, têm tipicamente atualizações em intervalos de meses, a Wikipédia é potencialmente atualizada em tempo real. O facto de haver liberdade na escolha de temas para os artigos, salvo as necessárias e raras restrições, significa uma maior diversidade de temáticas, capaz de agradar virtualmente aos utilizadores, havendo ainda a possibilidade de solicitar um artigo de interesse em particular à comunidade. Outro atributo que influencia positivamente a qualidade da informação é o nível de escolaridade dos seus utilizadores, em particular dos editores, onde segundo um estudo multilingue realizado pela United Nations University [68], que compreendeu 176.192 utilizadores em 22 idiomas diferentes, cerca de metade dos editores de conteúdo tem um grau de escolaridade correspondente ao Ensino Superior, com 4,4% com doutoramento e 18,7% com grau de mestre.

⁷<https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Fancruft>

O resultado do equilíbrio entre os fatores que contribuem positivamente e os que contribuem negativamente para a qualidade da informação disponibilizada pela Wikipédia é alvo de interesse, pelo valor intrínseco da informação. Em 2005 foi realizado um polémico estudo comparativo entre a Wikipédia e a Encyclopaedia Britannica [21], que revelou uma surpreendente semelhança quanto aos problemas de qualidade de ambas as enciclopédias, numa avaliação de pares, relativa a artigos relacionados com ciência. Das 42 avaliações, foram detetados apenas 8 erros graves, metade em cada uma das enciclopédias, tendo sido encontrados além destes outros erros menos graves, como omissões. Este estudo surpreendeu pela positiva quanto à qualidade da Wikipédia, e pela negativa, quanto à qualidade da Encyclopaedia Britannica, pelos erros detetados.

Dada a diversidade de conteúdos presentes na Wikipédia, surgiram diversos estudos sobre outras temáticas, como por exemplo sobre política norte-americana, num estudo de Brown, em 2011 [5]. Este estudo permitiu concluir que a informação se mostrou precisa, com os problemas a revelarem-se ao nível da cobertura, com falhas sobretudo em tópicos mais antigos ou menos proeminentes. Estes resultados permitiram ao autor concluir que de facto a Wikipédia não deve ser encarada como fonte completamente fiável para trabalhos rigorosos, mas é uma boa ferramenta para utilizadores sem acesso a outros recursos e que permite obter uma vista geral e rápida sobre uma temática com elevada fiabilidade.

Outros estudos continuaram a ser feitos sobre a qualidade da Wikipédia e, em 2015, Mesagri *et al.* [40] fizeram uma revisão sistemática com 110 artigos sobre o conteúdo da informação da Wikipédia, nomeadamente no que concerne à qualidade e à quantidade. O seu trabalho reuniu, assim, estudos que vão no sentido de afirmar a Wikipédia como fonte fiável de informação, mas outros que apontam o contrário, ainda que em termos gerais se tenha assumido como uma boa fonte de informação.

Corroborando em parte estas conclusões sobre a boa qualidade da informação presente na Wikipédia, um estudo levado a cabo em 2017 [65] evidenciou que a Wikipédia molda as pesquisas científicas, quer na contextualização das suas próprias pesquisas, quer no encaminhamento para artigos científicos da sua área de pesquisa. A Wikipédia afigura-se, assim, como uma importante fonte de disseminação de conhecimento, exercendo uma forte influência no conhecimento científico.

2.4 Mecanismos internos de qualidade

Tendo em conta as questões levantadas sobre a qualidade da informação presente na Wikipédia e a importância desta para a sua sobrevivência e crescimento enquanto fonte de informação, surgiu desde o início a necessidade de criar mecanismos internos para garantir níveis de qualidade aceitáveis. Existem três formas principais para assegurar os níveis de qualidade. Primeiro, surge a vantagem oferecida pela possibilidade de edição de conteúdo fácil e imediata por qualquer um; os próprios utilizadores são um garante de qualidade enquanto corretores de erros detetados aquando da utilização da informação. Em seguida temos o conjunto de políticas e de linhas orientadoras

próprias da Wikipédia⁸, criadas com a questão da qualidade em vista, que são usadas por equipas de voluntários organizados em departamentos específicos para analisar o material adicionado. Existem ainda *bots* que monitorizam de forma automática e contínua o conteúdo, procurando erros.

Derivados destas três formas gerais, foram criados mecanismos mais específicos para garantir os níveis de qualidade. Um desses mecanismos são os “critérios de notoriedade”, que definem se determinado tópico é merecedor de um artigo ou não. Dentro destes, a consideração mais geral é que “um tópico é presumido como notável se recebeu cobertura significativa de fontes reputadas e independentes do assunto tratado” [96]. Dada a enorme quantidade de conteúdos existentes, tornou-se necessário e prático dividir as páginas em grupos a serem monitorizados por grupos de utilizadores, humanos ou *bots*, designados de “Patrols”⁹. Esta monitorização é facilitada pelo facto de haver um registo de todas as edições feitas, *i.e.*, existe um histórico de cada página, que ajuda a compreender a evolução de um artigo ao longo do tempo, que também facilita o processo de monitorização de qualidade e correção, se necessário. As edições mais recentes, as maiores e as anónimas são também passíveis de serem identificadas em particular para facilitar o processo de monitorização. É ainda possível a cada utilizador criar a sua própria lista de artigos sobre os quais pretende manter vigilância ativa, que é atualizada de forma automática aquando de edições feitas aos mesmos.

2.4.0.1 Estrutura editorial

Da abordagem à garantia de qualidade, baseada em diferentes mecanismos, faz ainda parte a própria estrutura editorial da Wikipédia e os processos de controlo que a mesma encerra em si. Mantendo o espírito aberto e de voluntariado, a estrutura editorial é assim composta por voluntários, da qual qualquer um pode fazer parte, sujeitando-se a uma revisão paritária, sendo definida e auto-organizada pela própria comunidade. A administração é encabeçada pela Wikimedia Foundation (WMF) a quem a Wikipédia pertence [75], mas com pouca intervenção nas políticas desta, que são, na prática, entregues maioritariamente à comunidade de utilizadores.

No conjunto de utilizadores são diferenciados os meros leitores, dos editores – aqueles que colaboram na edição de conteúdo. Atualmente já se contam mais de 950 milhões de edições de conteúdo, numa média de quase 19 edições por página e um total de mais de 39 milhões de utilizadores registados [78]. Teoricamente todos os editores se situam na mesma escala hierárquica, no entanto, alguns deles têm privilégios e direitos de acesso que os permitem categorizar em seis grupos principais, conforme ilustrado no organograma 2.1, que mostra a hierarquia dos vários grupos dentro da Wikimedia.

O grupo dos “stewards” é um grupo global de utilizadores com acesso total à interface. Podem modificar todos os direitos locais e globais dos utilizadores, alterar o estado e o nome das contas globais, e aceder a todas as permissões disponíveis para administradores e burocratas. São eleitos pela comunidade global da Wikipédia.

⁸https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

⁹<https://en.wikipedia.org/wiki/Wikipedia:Patrols>

Os “burocratas” têm permissões para gerir os privilégios dos utilizadores, nomeadamente o de administrador. São utilizadores, normalmente administradores, escolhidos por consenso da comunidade após um pedido de aprovação.

Os “wikipedistas residentes” são editores dos projetos Wikimedia, que trabalham normalmente numa instituição cultural ou educativa, com o objetivo de incentivar e apoiar uma parceria entre a sua instituição e os projetos Wikimedia. A estes, é permitida uma compensação, económica, ou de outro tipo.

Por fim, temos os “educadores e estudantes”. O “Programa de Educação”¹⁰ ajuda na coordenação de educadores e estudantes de todo o mundo, contribuindo para a Wikipédia e outros projetos da Wikimedia, num ambiente acadêmico, com base em tópicos relacionados com o curso.

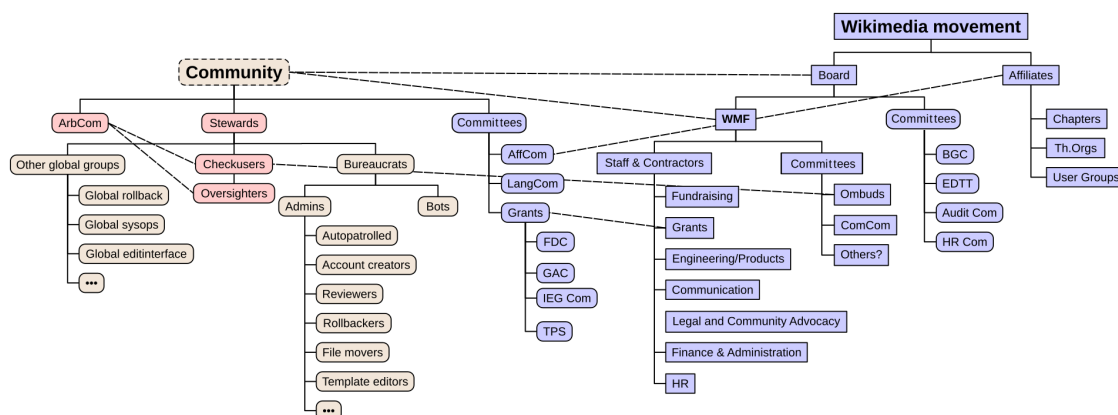


Figura 2.1: Hierarquia de grupos na Wikimedia. Fonte: <https://en.wikipedia.org/wiki/Wikipedia:Administration>

2.4.0.2 Controllo editorial

O primeiro tipo, refere-se à supervisão dos utilizadores. A grande quantidade de utilizadores, onde os bem-intencionados suplantam grandemente os mal-intencionados, e as características

¹⁰https://en.wikipedia.org/wiki/Wikipedia:Education_program

únicas de cada um trabalham em conjunto para um resultado comum, com ajuda das ferramentas disponibilizadas.

O segundo tipo, diz respeito à construção colaborativa de conhecimento. Ainda que, não dependa exclusiva ou maioritariamente de trabalho de especialistas, o grande número de editores garante à partida neutralidade e diferentes pontos de vista.

No terceiro tipo, temos a estrutura *wiki*, um princípio associado ao anterior. Salienta o facto de existir apenas uma mesma página para todos, sem a possibilidade dos editores que discordam conseguirem escrever artigos ou versões alternativas para expressar seus pontos de vista diferentes, havendo assim uma pressão para o consenso entre todos, e a almejada neutralidade. De salientar ainda, o facto de não haver entidades superiores de controlo do conteúdo, evitando manipulações motivadas por interesses secundários.

O quarto tipo, alude ao respeito pelas políticas e princípios. As regras, as políticas e princípios, são definidos tendo em conta um equilíbrio entre as boas e as más práticas, de modo a garantir a boa conduta e funcionamento por um lado e garantindo por outro que se mantém o potencial disruptivo, necessário para a evolução.

O quinto tipo, refere-se ao *ethos* baseado em consenso. Consenso significa que as preocupações e opiniões das minorias devem ser levadas em consideração na tentativa de obter uma decisão que reflita os valores da comunidade e que a maioria possa viver de uma maneira ou de outra. Mesmo no caso de disputa, o processo permanece o mesmo – até as decisões do Conselho de Arbitragem são baseadas em informações, consenso e transparência da comunidade.

Como sexto tipo, temos os processos de escalonamento e de resolução de disputas. Há mecanismos de escalonamento intrínsecos, como o facto de artigos mais sujeitos a vandalismo suscitarem uma vigilância mais apertada por parte dos utilizadores no sentido de o debelar. Há, ainda, mecanismos extrínsecos, criados pela comunidade, como a possibilidade de qualquer um poder solicitar a resolução de disputas, em fases progressivas conforme necessidade. No topo da cadeia, temos o Conselho de Arbitragem, para adotar uma resolução final, quando esta não seja possível por meios intermédios.

O sétimo tipo diz respeito ao *software* facilitador de monitorização e edição. Existem ferramentas de *software*, utilizadas pelos editores mais ativos, como o Huggle¹¹, para deteção automática de vandalismo em tempo real, entre outras ferramentas facilitadoras da identificação e correcção de problemas de qualidade. Existem, também, grupos de utilizadores dedicados à reversão do vandalismo, como o “Projeto AntiVandalismo”¹²;

O oitavo tipo, refere-se aos sistemas de bloqueio e proteção. Existem ferramentas para bloquear editores problemáticos e proteger páginas de editores de baixa qualidade, capazes de filtrar combinações de contas ou endereços *Internet Protocol* (IP). São utilizadas para impor bloqueios de curto ou longo prazo, e para bloquear páginas e impedir o vandalismo, conforme necessário.

¹¹<https://en.wikipedia.org/wiki/Wikipedia:Huggle>

¹²<https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Projetos/AntiVandalismo>

Por fim, temos a categorização da informação. Podem ser utilizadas etiquetas *inline* no texto para sinalizar declarações e citações individuais, ou artigos como um todo, para solicitar verificação ou citação e indicar a outros utilizadores que um facto ou apresentação não é suportado como está. Existem editores que procuram estas etiquetas nos artigos, para trabalhar neles.

2.4.0.3 Avaliação de artigos

As diversas versões da Wikipédia, de cada idioma, possuem geralmente um sistema de classificação de artigos [80], que não é, no entanto, normalizado entre elas. Por exemplo, na versão inglesa, este sistema é baseado em letras, à imagem da maioria das restantes versões, pretendendo indicar o quão completo um artigo se encontra, tendo em conta diferentes fatores. As avaliações de qualidade são efetuadas sobretudo pelos membros dos WikiProjects¹³ através de *tags*. Existem, no entanto, níveis de avaliação que são atribuídos por editores independentes. Estas avaliações permitem determinar a qualidade da informação de áreas específicas e priorizar os trabalhos de acordo com as expetativas. De referir que esta avaliação não possui carácter oficial, servindo sobretudo os interesses de cada *WikiProject* que a elabora. A Tabela 2.1 representa o esquema de avaliação utilizada pela *Wikipedia:Version 1.0 Editorial Team*¹⁴.

Para além desta classificação, pode ainda existir uma classificação da prioridade ou importância de um artigo, que reflete o nível de expectativa ou desejo de que determinado tópico seja retratado, e a escala varia geralmente entre “sem importância”, até “importância máxima”. Esta avaliação da importância é relativa a cada WikiProject.

2.5 Diferenças entre idiomas

Segundo o Ethnologue [12], existem atualmente 7.139 idiomas vivos no mundo inteiro. Tal diversidade pode levantar, naturalmente, questões sobre a presença das mesmas na *web*. Em 2009, Pimienta *et al.* [48] descreveram uma investigação levada a cabo de 1996 até 2008, pela UNESCO através da FUNREDES e Union Latine sobre a diversidade linguística na *web* que revelou uma grande discrepância na presença de idiomas no ciberespaço. Recorrendo principalmente a um método baseado em motores de busca, é comparada a presença do idioma inglês com outros seis idiomas principais. Os resultados divulgados mostram uma clara dominância de páginas *web* em inglês, seguida do alemão, ainda que este nunca tenha chegado sequer aos 18%; abaixo deste, surgem o espanhol e o francês, seguido do italiano e português. Dentro dos idiomas estudados, surge em último lugar o romeno, que se manteve sempre abaixo de 1%.

Baseada na linguagem como meio principal de comunicação, a Wikipédia assume-se como um indicador do multilinguismo *online* no conjunto dos vários idiomas presentes. A WMF tem definidas políticas para propor novos idiomas, criadas pelo seu “Language committee” [76], que é também o responsável pelo processamento das propostas e ainda por projetos associados ao assunto. Para a proposta ser elegível deve propor um idioma ainda não existente na Wikimedia,

¹³<https://en.wikipedia.org/wiki/Wikipedia:WikiProject>

¹⁴https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team

Tabela 2.1: Esquema de avaliação de artigos da Wikipédia

Classe	Avaliação	Alvo
FA	Artigo com nota máxima, avaliado exaustivamente por avaliadores independentes. É um artigo com boa prosa, compreensivo, com boa investigação subjacente, neutro no ponto de vista, estável, segue as normas de estilo, tem elementos multimédia apropriados e tem uma extensão adequada ao seu conteúdo.	Profissionais, utilizadores em busca de conteúdo verdadeiramente enciclopédico.
FL	Artigo identificado como sendo de excelente qualidade e passível de fazer parte de listas destacadas. Possui boa prosa, tem uma secção introdutória, é completo no seu âmbito de definição, está bem estruturado, cumpre as normas de estilo, tem apelo visual e estabilidade.	Profissionais em busca de informação fidedigna.
A	Artigo analisado por avaliadores imparciais. É bem organizado e essencialmente completo, bem escrito, claro, com uma extensão adequada, bem estruturado e bem referenciado.	Útil para leitores não especializados na matéria, mas em busca de informações completas.
GA	Artigo analisado por um ou mais avaliadores imparciais. Está bem escrito, está referenciado, sem pesquisas inéditas, tem uma cobertura do tema abrangente, é neutro no ponto de vista, estável, e ilustrado quando conveniente.	Para praticamente todos os leitores, com qualidade próxima de uma enciclopédia profissional.
B	Artigo analisado por um ou mais avaliadores imparciais. Está bem referenciado, cobre razoavelmente o tópico, sem omissões e erros óbvios, tem uma estrutura definida, está razoavelmente bem escrito, contém elementos multimédia apropriados e o seu conteúdo é compreensível.	Leitores em busca de informação praticamente completa, ainda que o conteúdo possa não ser adequado a estudantes.
C	Artigo onde falta conteúdo importante ou que contém muito conteúdo irrelevante. Tem referências para fontes credíveis, está estruturado, seguindo normas de estilo, mas com falta de alguns dos critérios necessários para a classe B.	Direcionado a leitores casuais, que não procurem uma imagem completa ou mesmo moderada do assunto.
Start	Artigo incompleto, que ainda está em fase de desenvolvimento. Pode não conter referências de fontes fiáveis, a prosa pode não ser de elevada qualidade, mas deve satisfazer as políticas gerais da Wikipédia.	Para leitores em busca de uma visão geral, não aprofundada de um assunto.
Stub	Descrição básica do tópico. Pode não estar bem redigido e ter problemas no conteúdo em si, é geralmente muito curto e corre o risco de deixar de ser considerado artigo.	Por apresentar conteúdo pouco significativo, é útil para quem procura uma definição ao estilo de um dicionário ou pouco mais.

Fonte: https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

deve ter um código ISO 639 1-3¹⁵ válido, estando maioritariamente excluídos dialetos regionais e formas diferentes do mesmo, e deve possuir um número suficiente de utilizadores fluentes para formar uma comunidade viável de colaboradores e público interessado no seu conteúdo. Para a aprovação final, é ainda requisito que exista um projeto-teste na Wikimedia e que exista um esforço contínuo para traduzir a interface da Wikimedia nesse idioma. Existem atualmente 266 propostas de novos idiomas a adicionar à Wikipédia¹⁶.

¹⁵https://en.wikipedia.org/wiki/ISO_639

¹⁶https://meta.wikimedia.org/wiki/Requests_for_new_languages

2.5.1 Artigos e utilizadores da Wikipédia

Usando os dados estatísticos disponibilizados pela Wikipédia [77] à data de Maio de 2021, foi criada a Tabela 2.2, onde constam o número de artigos, edições, administradores, utilizadores, utilizadores ativos e quantidade de imagens para cada versão da Wikipédia com mais de 1 milhão de artigos, ordenada por ordem decrescente do número de artigos.

Tabela 2.2: Estatísticas para versões da Wikipédia com mais de 1 milhão de artigos

	Artigos	Edições	Admin.	Util.	Util. ativos	Imagens
Inglês	6.303.060	1.019.830.553	1.097	41.588.077	137.410	892.724
Cebuano	5.739.406	31.488.950	6	79.002	162	0
Sueco	3.173.043	49.187.385	63	776.319	2.564	0
Alemão	2.579.016	210.805.034	187	3.707.993	19.915	129.797
Francês	2.331.163	182.692.698	156	4.102.244	21.855	64.641
Neerlandês	2.056.164	58.848.118	35	1.143.068	4.323	21
Russo	1.725.851	114.065.292	79	2.981.113	11.557	230.861
Italiano	1.694.970	120.508.728	114	2.125.584	9.687	141.287
Espanhol	1.686.066	135.284.609	67	6.226.574	17.205	0
Polaco	1.475.293	63.194.765	102	1.104.183	4.729	260
Árabe egípcio	1.291.122	5.609.669	6	154.386	205	1.454
Japonês	1.269.739	83.437.629	41	1.789.706	15.570	54.737
Waray	1.265.326	6.233.912	3	48.308	74	42
Vietnamita	1.264.057	64.886.869	20	803.046	2.401	22.719
Chinês	1.198.452	65.438.153	78	3.088.085	8.438	55.950
Árabe	1.117.135	53.808.359	27	2.079.062	5.569	45.946
Ucraniano	1.093.256	31.852.374	45	560.294	3.458	111.224
Português	1.066.663	61.033.778	71	2.590.868	10.280	57.525

Fonte: https://en.wikipedia.org/wiki/List_of_Wikipedias

O primeiro dado que se pode extrair da tabela é o facto de existirem 18 idiomas na Wikipédia com mais de um milhão de artigos. Em seguida, confirma-se a superioridade do inglês quanto à quantidade de artigos disponíveis, seguido pelo cebuano, uma língua austronésia falada por cerca de 18,5 milhões de pessoas nas Filipinas [63]. Isto justifica-se pelo facto de existir um *bot*, o Lsj-bot¹⁷ que conta já com mais de 17 milhões de artigos criados, sendo responsável pela maioria dos artigos escritos em cebuano, em sueco e em samarês [73], o que também explica o facto do sueco aparecer em terceiro lugar, seguido só então pelo alemão. Relativamente ao número de edições, o inglês mantém o domínio. O cebuano e o sueco deixam de ocupar os lugares cimeiros, o que também está em consonância com o facto dos artigos serem escritos por um *bot*, que neste caso, não realiza edições posteriores à criação dos artigos. Assim, a seguir ao inglês, ainda que com grande diferença, surge o alemão, seguido pelo francês. Quanto ao número de administradores, os dados estão em consonância com os números de edições. Analisando o número de utilizadores, verifica-se que o inglês mantém um domínio acentuado, surgindo o espanhol em segundo lugar. De notar, a diferença significativa entre o número total de utilizadores e de utilizadores ativos – utilizadores registados que exerceram alguma ação na Wikipédia nos últimos 30 dias. Por fim, quanto ao número de imagens, reforça-se o domínio do inglês, que é agora seguida do russo em

¹⁷<https://sv.wikipedia.org/wiki/Anv%C3%A4ndare:Lsjbot>

segundo lugar. O número de imagens corresponde àquelas carregadas localmente, sendo que existem algumas versões que não usam imagens locais, dependendo completamente do Wikimedia Commons¹⁸, explicando assim os valores de 0 presentes. De modo geral, há a indicação de uma quantidade considerável de conteúdo multimédia, além do textual, o principal da Wikipédia.

Em 2009, Dijk [69] abordou a questão das edições da Wikipédia em idiomas minoritárias e formas de as mensurar para comparar. Ele refere a obsessão com o número de artigos em cada Wikipédia e a comparação com outros idiomas. Ele concluiu ser difícil atribuir os fatores que de facto contribuem para o crescimento de cada versão da Wikipédia, mas deu especial relevo à quantidade de falantes de cada idioma, uma vez que representam o potencial de editores de artigos desse idioma. Isto, no entanto, poderá não corresponder à realidade, pois Matei em 2017 [39] conclui num estudo com dados de edições da primeira década de existência da Wikipédia, que 77% dos artigos foram criados por apenas 1% dos editores, o que levanta até problemas sobre o espírito colaborativo da mesma, estando, no entanto, este assunto fora do âmbito desta dissertação. Nas suas conclusões, Dijk refere ainda a importância da atitude das pessoas perante projetos como a Wikipédia, apontando este como o principal fator para o crescimento dos idiomas latinos na Wikipédia, e termina com a importância da colaboração de instituições ligadas às questões linguísticas para o desenvolvimento de conteúdos, sobretudo em idiomas minoritários.

Mais tarde, em 2011, surge um estudo de Hale [24] sobre o papel dos editores multilingues na edição das várias versões da Wikipédia, enquanto fomentadores de desenvolvimento dos diversos idiomas dentro desta. É também referida a importância do sistema de inter-línguas da Wikidata [74], um sistema lançado em 2012 pela WMF, que em conjunto com os *links* interlinguísticos¹⁹ proporciona uma solução centralizada, baseada numa base de dados colaborativa, que permite conetar o mesmo conceito em diversas versões da Wikipédia e até mesmo entre outros projetos da Wikimedia. São armazenados essencialmente itens, cada um com uma etiqueta, uma descrição e uma lista de nomes alternativos, interligando os itens e os seus dados. Hale concluiu que a maior parte dos editores atua apenas num único idioma, havendo 15% que o fazem em idiomas diferentes, com a particularidade destes serem, por norma, mais ativos que os restantes.

Dijk [69] menciona o facto de haver artigos na Wikipédia com pouco ou nenhum detalhe, uma espécie de “pseudo-artigos”, nomeadamente artigos extremamente curtos, geralmente com uma frase, semelhantes a dicionários, “stubs” geográficos – pequenos artigos sobre cidades e vilarejos que fornecem apenas informações geográficas limitadas e artigos de tradução – que geralmente fornecem poucos detalhes. Assim sendo, o de número de artigos pode revelar uma verdade enganadora sobre a qualidade das versões da Wikipédia.

2.5.2 Qualidade do conteúdo

Num artigo publicado em 2009, Filatova [15] procura descrever o multilinguismo da Wikipédia através de uma *framework* criada para o efeito e utilizando apenas o texto dos artigos. É referido que artigos sobre a mesma coisa, diferem muito entre versões, sobretudo ao nível da

¹⁸https://en.wikipedia.org/wiki/Wikimedia_Commons

¹⁹https://pt.wikipedia.org/wiki/Ajuda:Guia_de_edi%C3%A7%C3%A3o/Interwikis

quantidade de informação coberta em cada versão mas também pela informação que os editores escolhem para ser coberta no artigo em si, afetando diretamente a qualidade deste.

Domingues e Teixeira Lopes [9], levaram a cabo um estudo comparativo sobre a qualidade de artigos relacionados com medicina nas versões portuguesa e inglesa da Wikipédia em 2019. Os autores concluíram haver diferenças significativas entre as duas versões na grande maioria das métricas avaliadas. Os resultados sugerem que os artigos em inglês demonstram um esforço maior na organização do conteúdo, reutilização de informação e utilização de citações. A conclusão geral foi que há, de facto, uma grande variação na qualidade de artigos relacionados com a saúde, a favor da versão inglesa.

Apesar da escassa investigação disponível sobre as diferenças na qualidade entre as diferentes versões da Wikipédia de acordo com o idioma, parece haver uma relação direta entre os aspetos quantitativos e qualitativos da informação disponível. Assumindo que idiomas com menor expressão quantitativa na Wikipédia traduzem informação de menor qualidade, e dada a importância da Wikipédia enquanto fonte de informação, estamos perante um problema de desigualdade que mereceu a atenção da UNESCO [67]. Esta, reconhece que a informação presente no ciberespaço é um fator muito importante para o desenvolvimento da humanidade, na medida em que é um modo primário de partilha de informação e conhecimento. Tem, assim, envidado esforços no sentido de combater as desigualdades no acesso à informação *online*, tanto a nível de disponibilidade como de possibilidade de aceder à mesma.

2.6 Informação de saúde na Wikipédia

Reconhecida a importância do acesso à informação por todos, o facto de haver ainda grandes desigualdades no mesmo, e o potencial da *web* para debelar esse problema, a Wikipédia assume um papel de destaque por todas as suas características que fazem de si aquilo em que se tornou.

Em 2015, Heilman *et al.* [26] calcularam que no final de 2013, a Wikipédia tinha mais de 155 mil artigos médicos, distribuídos por 255 idiomas e cerca de 19% destes artigos eram em inglês, correspondentes a mais de 29 mil artigos, assumindo assim a liderança com uma grande margem em relação ao alemão, em segundo lugar, com menos de oito mil artigos. Havia 31 idiomas que não continham qualquer artigo médico.

Laurent e Vickers analisaram, em 2009 [31], a importância da Wikipédia enquanto fonte de informação *online*, através da sua relevância nos resultados em motores de busca com termos médicos e concluíram que esta não só está presente no top 10 dos resultados de busca em mais de 70% dos casos, como ainda demonstrou ter elevada qualidade nos artigos encontrados, o que, de forma geral, confirmou a proeminência da Wikipédia enquanto fonte de informação *online* para a saúde, comparativamente a outras fontes *online*. Num artigo publicado em 2017, Shafee *et al.* [56] analisaram a evolução dos conteúdos da área médica na Wikipédia e concluíram que a qualidade tem vindo a melhorar desde a sua criação, para o qual contribuem as suas colaborações com institutos médicos e universidades, mas rematam que ainda há espaço para melhoria e para esta apelam à contribuição da comunidade médica.

As características da Wikipédia enquanto fonte de informação em saúde tornam-na apelativa aos estudantes da área da saúde. É o caso dos estudantes de enfermagem, ainda que segundo Haigh [23] seja vista como uma fonte de informação suspeita, mas que no seu estudo, em 2011, concluiu ser apropriada para estudantes de enfermagem. O autor deu especial ênfase ao seu papel como ponto de partida para a procura de informação. Outro estudo, realizado em 2015, por Prasannan *et al.* [50] conclui também que a Wikipédia é utilizada por estudantes de medicina, sendo a fonte mais popular a seguir à UpToDate²⁰. O mesmo estudo revelou que apenas 20% dos estudantes utilizavam os livros como fonte de informação principal. Outro estudo realizado em 2017 [62], também em estudantes de medicina, concluiu que a Wikipédia era a fonte de informação *online* mais utilizada por estes durante o seu estágio em cirurgia, e mais de metade deles não recorreu a livros. Cientes da importância que a Wikipédia assumiu como fonte de informação para os estudantes da área da saúde, Azzam *et al.* [2] propuseram durante dois anos um curso de edição da Wikipédia a estudantes de medicina, com o objetivo – alcançado – de melhorar a informação médica de artigos da Wikipédia, e apresentam como sugestão, um maior envolvimento de outras instituições de ensino em projetos idênticos, visando entender a influência dos estudantes de medicina na Wikipédia e vice-versa.

Investigadores do Botswana [46] estudaram as necessidades de informação clínica dos seus profissionais de saúde e a sua percepção sobre a Wikipédia enquanto ferramenta para as satisfazer. É referido pelos investigadores que os profissionais de saúde do Botswana enfrentam sérias dificuldades no acesso a informação de saúde, que poderia ser colmatada pela utilização da Wikipédia. As fontes de informação utilizadas estavam desatualizadas, não eram práticas em situações de maior pressão e continham discrepâncias na fiabilidade da informação, revelando-se claramente insuficientes e pouco fiáveis, com consequências negativas ao nível da confiança transmitida aos utentes. A opinião dos profissionais quanto à Wikipédia manifestou-se heterogênea, levantou questões de necessidade de validação prévia pelas entidades oficiais de saúde e levantou ainda problemas com a disponibilidade *offline*, importante em regiões sem acesso à Internet.

2.6.1 WikiProject Medicine

Com a dimensão e relevância que a área da saúde ganhou dentro da Wikipédia, assistiu-se à criação de um WikiProject nesta área – o WikiProject Medicine [95], existente em várias versões da Wikipédia, incluindo na portuguesa²¹. Trata-se de um projeto iniciado em 2004 com o objetivo de contribuir para uma melhoria dos artigos médicos, sendo um projeto integrante da Wiki Project Med Foundation²². Conta atualmente com 354 membros ativos. Não é necessário qualquer requisito formativo aos seus membros, mas a maioria deles são médicos, estudantes de medicina, enfermeiros, cientistas, mas também utentes e leigos com interesse específico em determinados tópicos médicos [25].

²⁰<https://www.wolterskluwer.com/en/solutions/uptodate>

²¹<https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Projetos/Sa%C3%BAd>

²²https://meta.wikimedia.org/wiki/Wiki_Project_Med

Neste projeto foram definidas políticas adicionais às da Wikipédia em geral, e ao longo do tempo foram reunindo um conjunto de ferramentas para ajudar os seus membros a atingir os objetivos preconizados. Este projeto realiza trabalhos de parceria para investigação e publicação científica, programas educacionais na área médica e ainda projetos para melhorar o acesso à informação, como a *app* móvel “Medical Wikipedia”²³ ou o “Internet-in-a-box”²⁴ para acesso a informação *offline*. Há ainda iniciativas para tradução de artigos em diversas idiomas, ajudando assim a eliminar as assimetrias que se verificam [79].

Uma das tarefas desempenhadas com grande impacto na qualidade da informação é a avaliação da qualidade de artigos médicos, de acordo com critérios definidos e baseados nos critérios de avaliação definidos para a Wikipédia em geral, somando atualmente mais de 55 mil artigos avaliados [89].

²³https://meta.wikimedia.org/wiki/Wiki_Project_Med/App

²⁴<https://meta.wikimedia.org/wiki/Internet-in-a-Box>

Capítulo 3

Qualidade da informação existente na Wikipédia

Neste capítulo, será descrita a revisão de literatura relativa à qualidade da informação da Wikipédia, com especial enfoque nas métricas de avaliação da qualidade de informação. Nesse sentido, depois de uma introdução ao tema, é abordado o conceito de qualidade da informação. Posteriormente são abordadas a qualidade da informação, em geral, e em específico na saúde. O capítulo termina com uma análise aos trabalhos relativos aos dois tipos de qualidade.

3.1 Introdução

A Figura 3.1 é uma representação de uma hipotética impressão atual da Wikipédia na sua versão em inglês, sem imagens, usando como referência volumes da Encyclopaedia Britannica e fornece uma boa perspetiva da dimensão da Wikipédia. Estão representados 2.657 volumes em 14 pilhas. A mesma representação para as versões da Wikipédia de todos os idiomas é composta por 20.120 volumes em 101 pilhas¹. Analisar toda a informação aí contida, e em particular a qualidade dessa informação, revela-se portanto um enorme desafio. Para superar esse desafio, a investigação centra-se sobretudo em amostras, sendo que a validade dos dados dependerá do tamanho da amostra, o que levanta de igual forma desafios na recolha e tratamento dos dados. Quando os investigadores pretendem analisar amostras de tamanho considerável, são levados a métodos mais ou menos automatizados que facilitem a tarefa tornando-a possível.

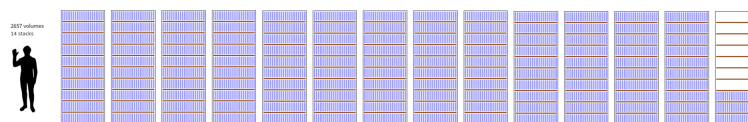


Figura 3.1: Estimativa do tamanho de uma versão impressa da Wikipédia. Fonte: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

¹https://meta.wikimedia.org/wiki/Wikipedia_as_books

De entre as abordagens à avaliação da qualidade da Wikipédia, podemos distinguir aquelas baseadas em *machine learning*, sobretudo em trabalhos mais recentes, e aquelas baseadas em medidas e métricas mais concretas e “explicáveis”. Em seguida serão focados os trabalhos baseados no segundo tipo de abordagem referido, sub-dividindo-o em avaliação da qualidade em geral e, posteriormente, em particular, na área da saúde.

3.2 O conceito de qualidade da informação

Não há uma definição única para o conceito de qualidade. Vivek Nanda [43], categorizou, de forma geral, o conceito em dois aspetos – satisfação de especificações e satisfação de consumidores. Tal como para a qualidade, também não existe uma única definição para a informação, que varia com o contexto. Numa tentativa de conseguir captar o conceito o o mais independentemente possível do contexto, Madden [36] define-a como “um estímulo originado num sistema que afeta a interpretação por outro sistema da relação do segundo sistema com o primeiro ou do relacionamento que os dois sistemas compartilham com um determinado ambiente”. O autor admite, no entanto, que possivelmente nunca haverá uma definição consensual para a informação.

Com a subjetividade dos dois conceitos anteriores, o conceito de qualidade da informação é, como seria de esperar, também ele subjetivo, abrangendo múltiplas dimensões, como sugerido por Miller [41]. O autor identificou 10 dimensões para a qualidade de informação: relevância, precisão, intemporalidade, completude, coerência, formato, acessibilidade, compatibilidade, segurança e validade. Mais uma vez, é referido que o valor da informação reside no consumidor e nas suas perceções da mesma. Lee *et al.* [32] definiram, por seu lado, quatro dimensões no conceito de qualidade de informação: intrínsecas, contextuais, representacionais e de acessibilidade.

Dadas as diferentes dimensões que a qualidade da informação pode adotar, e sobretudo a dependência do contexto, os trabalhos de avaliação da qualidade da informação da Wikipédia são, aqui, divididos nos gerais e naqueles em contexto específico da saúde, e apresentados nas secções seguintes.

3.3 Qualidade em geral

Quando se aborda a avaliação da qualidade da informação na Wikipédia, o trabalho de Stvilia *et al.* [58] é incontornável. Utilizando dois conjuntos de artigos da Wikipédia inglesa, um deles com artigos aleatórios e outro apenas com “Featured Articles”, os autores extraíram o histórico de edições e páginas de discussão destes. Partindo dos tipos de problemas relatados pelos utilizadores nas páginas de discussão e das “Featured Article Removal Candidates”, e adaptando métricas já existentes para avaliação da qualidade na *web*, definiram sete métricas: *Authority*, *Completeness*, *Informativeness*, *Complexity*, *Consistency*, *Volatility* e *Currency*.

Uma vez que as métricas utilizadas por Stvilia *et al.* implicam a recolha e processamento de várias medidas de primeiro nível, extraídas dos dados, autores como Blumenstock [4] apostam em medidas mais simples. Assim, em 2008, Blumenstock, utilizou apenas o número de palavras dos

artigos. A justificação foi a facilidade de medir, de obter os dados, ser facilmente interpretável e que referiu como tendo bom desempenho. Esta medida foi validada recorrendo aos “Featured Articles” e foram avaliados mais de 11.000 artigos da Wikipédia inglesa. Utilizando um limiar de 2.000 palavras, conseguiu uma precisão calculada superior a 96% na avaliação binária dos artigos em “featured” e “random”.

Cinco anos após o trabalho de Stvilia, Wu *et al.* [98], publicaram um trabalho onde estabeleceram um mapeamento entre as métricas de Stvilia e os próprios padrões de qualidade da Wikipédia. Utilizaram um total de 28 métricas divididas em quatro grupos: linguísticas, estruturais, históricas e reputacionais, como ilustrado na Figura 3.2. Estas métricas foram posteriormente utilizadas numa rede neuronal, avaliando 500 artigos de cada categoria da classificação feita previamente pela Wikipédia, num total de 4.500 artigos. Os autores concluíram que nos artigos com classificação inferior as métricas linguísticas assumiam mais importância, enquanto as estruturais e históricas eram mais importantes em artigos com classificações superiores, restando as métricas reputacionais que manifestaram pouca importância geral. Como conclusão geral, a métrica mais relevante para a qualidade foi a completude dos artigos. Uma conclusão importante foi de que a reputação dos editores não se manifestou relevante.

Em 2011, Xu e Luo [99] basearam-se nos rácios de oito características lexicais dos textos, nomeadamente: rácio de palavras por frase; rácio de palavras diferentes por frase; rácio de substantivos nas palavras; rácio de verbos nas palavras e rácio de verbos relacionais. Estes dados foram posteriormente aplicados numa árvore de decisão. Avaliaram 100 artigos com avaliação “Featured Articles” e mais 100 artigos com avaliação “Start”, num total de 200 artigos, conseguindo uma precisão geral superior a 83%.

Baseados na proposição de que, artigos com baixa qualidade têm maior probabilidade de serem editados, Suzuki e Yoshikawa [61] propuseram uma avaliação da qualidade baseada no rácio de sobrevivência de um texto às edições, onde um período de sobrevivência longo significará um texto de melhor qualidade. Por outro lado, associaram um *score* de qualidade do editor desse texto, de modo a diminuir a importância de edições feitas por vândalos. Foi utilizada a versão japonesa da Wikipédia e foram avaliados 484.146 artigos. Comprovaram, por um lado, que o *score* de qualidade convergia, e por outro, que o seu método conseguia avaliar eficazmente os *scores* de qualidade do texto.

Li *et al.* [34] propuseram, em 2015, uma solução também baseada na relação entre os artigos e os editores mas com mais passos intermédios, tendo criado diferentes modelos, desde um mais básico baseado no PageRank até modelos mais complicados. Avaliaram cerca de 50.000 artigos e as suas conclusões sugerem métodos de avaliação que combinem a rede de artigo-editor com uma avaliação manual da qualidade.

No mesmo ano, de la Robertie *et al.* [7] utilizaram também a relação entre artigos e editores, baseando-se no princípio do Reforço Mútuo, com uma distinção entre autores, editores e co-editores. Foi utilizada análise de grafos e das relações entre estes em conjunto com o algoritmo PageRank para análise da rede de grafos. Para validação de resultados foi utilizada a pré-avaliação dos artigos pelos respetivos WikiProjects de mais de 20.000 artigos.

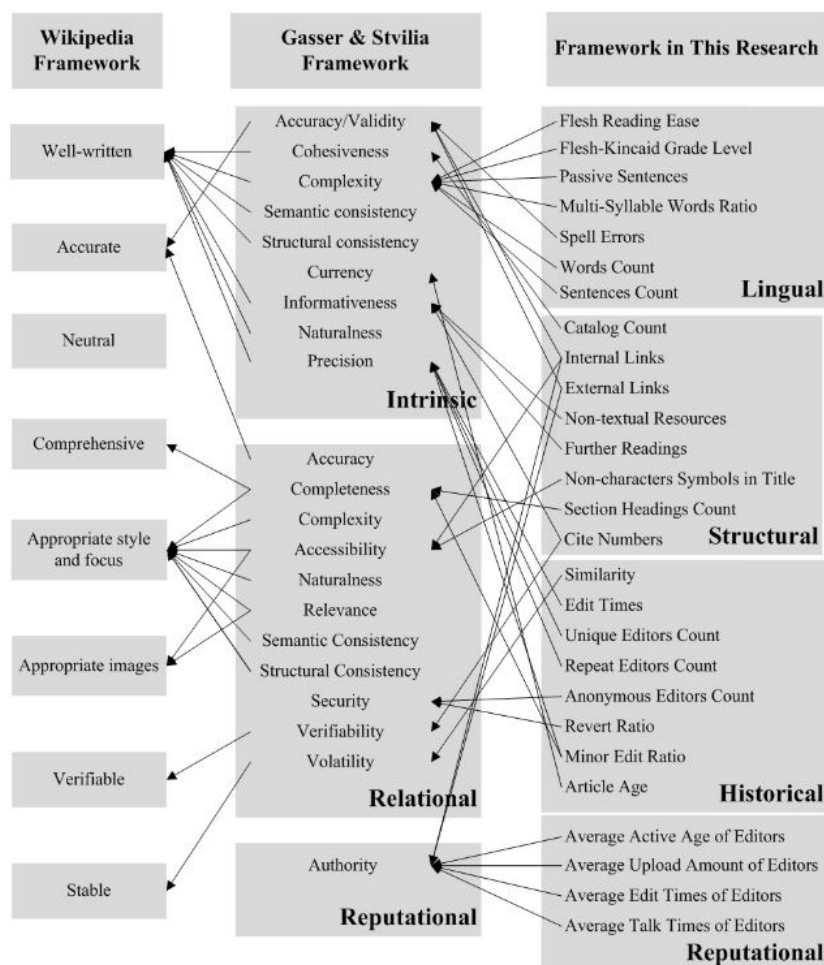


Figura 3.2: Mapeamento entre três modelos de avaliação. Fonte: Wu *et al.* [98]

Já em 2019, surge uma abordagem que se diferencia das anteriores por se basear apenas no histórico de edição dos artigos. Assim, Marrese-Taylor *et al.* [37], recorrendo a medidas mais tradicionais, baseiam-se na edições dos artigos, mas acrescentam no entanto, e como novidade, a descrição que acompanha cada edição. No seu modelo, é calculada a diferença entre duas versões do mesmo artigo, após uma edição, e é incorporada a descrição que o editor faz das suas edições. Foram utilizados artigos das versões inglesa e alemã da Wikipédia, de onde foram extraídos os artigos com mais edições, num total de cerca de 30.000 artigos.

Por fim, noutra proposta recente, de 2019, Khan *et al.* [27] propõem uma abordagem que recorre a medidas comuns, nomeadamente o número de palavras do artigo, número de edições e idade do artigo. Foram avaliados 100 artigos recentes – de final de 2018 – da Wikipédia em inglês recorrendo à avaliação feita pela Wikipédia sobre a qualidade dos artigos. Os resultados demonstraram que as três medidas são reveladoras da qualidade dos artigos segundo a pré-avaliação da Wikipédia. O escasso tamanho do *dataset* levanta no entanto dúvidas quanto à validade das conclusões.

3.4 Qualidade em saúde

Dentro dos trabalhos de avaliação da qualidade da informação da Wikipédia é possível encontrar alguns que se focam na saúde e medicina. A esfera da saúde está dividida em várias áreas, como é o caso da oncologia, sobre a qual incidiu uma pesquisa de Rajagopalan *et al.* [52]. Este trabalho incidiu sobre a qualidade da Wikipédia em relação a bases de dados profissionais específicas do ponto de vista do utente e sua família. Esta avaliação da qualidade foi feita de forma manual, sendo que os critérios se prenderam com a precisão, aspetos controversos, inteligibilidade, referências, revisões/edições e informação nova. Foram usados os artigos sobre 10 tipos de cancro da Wikipédia em comparação com os respetivos artigos na “National Cancer Institute’s Physician Data Query”. Os autores concluíram que, a nível de qualidade, ambas as fontes se encontravam no mesmo patamar, com a Wikipédia a perder apenas ao nível da inteligibilidade.

Outra das áreas alvo de estudo foi a da nefrologia, onde um trabalho de 2013, de Thomas [64] avaliou a qualidade da Wikipédia, utilizando como métricas a compreensibilidade, a confiança e inteligibilidade. Foram utilizados os códigos de diagnóstico de nefrologia do “International Statistical Classification of Diseases and Related problems, 10th Edition” para pesquisar artigos na versão inglesa da Wikipédia. A compreensibilidade foi avaliada pela proporção de códigos que tinham artigos correspondentes. A confiança foi avaliada pelo número de referências do artigo e pela proporção de referências de fontes cientificamente comprovadas. A inteligibilidade foi avaliada usando três índices: Flesch-Kincaid, Índice de Inteligibilidade Automatizada e Flesch reading ease. Este foi um trabalho de avaliação semi-automática. Os autores classificaram a Wikipédia como bastante compreensiva, fiável e com um nível de escrita correspondente a um nível de formação superior.

Em 2014, Conti *et al.* [6] propuseram-se a avaliar o grau de maturidade de mais de 2.400 artigos médicos da Wikipédia inglesa, recorrendo ao “Analytic Hierarchy Process” – um método multi-critério de apoio à tomada de decisões complexas, baseado em matemática e psicologia, desenvolvido na década de 1970. As medidas utilizadas prenderam-se com critérios linguísticos, estruturais, de histórico e reputacionais. Aquelas ligadas aos critérios linguísticos foram: índices Flesch reading ease e Flesch-Kincaid; contagem de palavras e contagem de frases; rácio de palavras e palavras com várias sílabas; rácio de erros ortográficos e quantidade de palavras. Quanto às ligadas às estruturais, foram: número de categorias; *links* internos e externos; recursos multimédia; número de referências adicionais; símbolos no título; número de títulos de secção e número de citações. Aquelas referentes ao histórico foram: número de edições; número de editores; número de editores que editaram o artigo mais de uma vez; número de editores anónimos; fração de edições marcadas como menores pelo editor; idade do artigo e número de edições. Por fim, as ligadas a critérios reputacionais foram: idade média dos editores; quantidade média de *upload* de editores; tempo médio de edição e tempo médio de conversação dos editores. As suas conclusões principais foram que a quantidade nem sempre corresponde à qualidade, sobretudo no que se refere ao conceito de maturidade estudado.

Outra das áreas estudadas foi a da neurocirurgia por Modiri *et al.* [42] em 2018. A informação

disponível na Wikipédia foi avaliada de forma diferenciada ao nível da qualidade e inteligibilidade. Para avaliação da inteligibilidade foram utilizadas cinco escalas diferentes: “Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, Simple Measure of Gobbledygook Grade level e Coleman-Liau Index”. Para avaliar a qualidade foram utilizados o “Center for Disease Control Clear Communication Index”² e a DISCERN³. Foram avaliados 55 artigos da área de neurocirurgia da versão inglesa da Wikipédia. No mesmo ano, e também recorrendo à escala DISCERN e à escala de Flesch-Kincaid, Simpson e Malicka [57] publicaram um trabalho semelhante, mas ligado à perda de audição. Os autores de ambos os estudos identificaram algumas falhas de qualidade, mas sobretudo, de inteligibilidade na informação disponibilizada pela Wikipédia, em particular na versão anglófona.

Recentemente, London *et al.* [35] estudaram, em 2019, a qualidade da Wikipédia na área da anatomia musculoesquelética em comparação com a informação presente no Atlas de Anatomia Grant. Para esta comparação, e em contraste com os trabalhos anteriores, realizaram uma análise completamente manual, avaliando a precisão e completude de 101 artigos sobre músculos. Mais uma vez as conclusões apontam para uma qualidade satisfatória da Wikipédia, apontando como principal ponto negativo a validade científica das referências utilizadas.

Dentro também da área da anatomia, mas em geral, e no mesmo ano, Suwannakhan *et al.* [60] propuseram avaliar a qualidade e inteligibilidade da Wikipédia. Esta análise foi feita de modo semi-automatizado, recorrendo à ferramenta da Wikimedia X-tools em associação com a DISCERN. Para avaliar a qualidade, foi utilizada a data de criação do artigo, o historial de revisões e o tamanho do artigo. A inteligibilidade foi avaliada com recurso aos indicadores “Flesch-Kincaid Grade Level Readability Test” e “Flesch Reading Ease Readability Formula”. Recorreram ainda à avaliação manual de especialistas da área. Foram avaliados 40 artigos e os autores apontaram mais uma vez limitações da Wikipédia a nível da inteligibilidade dos artigos.

Por fim, e ainda em 2019, Domingues e Teixeira Lopes [9] compararam a qualidade da versão portuguesa com a versão anglófona da Wikipédia em artigos relacionados com medicina. Para tal, recorreram a métricas definidas por Stvilia *et al.* [58], mas também a medidas mais específicas, como o número de *templates* de medicina, número de *infoboxes* de medicina e número de citações. Os resultados mostraram uma qualidade superior da Wikipédia inglesa.

3.5 Análise comparativa

Dos trabalhos analisados, gerais e de saúde, é possível extrair alguns padrões nas práticas utilizadas, resultados conseguidos e nas medidas e métricas utilizadas. São evidentes algumas diferenças entre os trabalhos que envolvem a avaliação da qualidade da Wikipédia em geral e daqueles direcionados apenas para área da saúde e medicina, daí se ter optado pela distinção dos dois tipos de trabalho.

²<https://www.cdc.gov/ccindex/index.html>

³http://www.discern.org.uk/discern_instrument.php

A principal diferença evidenciada entre os dois tipos de trabalhos é a avaliação, que é sempre automática nos estudos da qualidade em geral, enquanto que naqueles orientados para a saúde, envolve quase sempre algum tipo de análise manual, ou até mesmo completamente manual. Associada a esta, verifica-se outra diferença, o tamanho dos *datasets*. Utilizando métodos automáticos, os trabalhos conseguem analisar uma maior quantidade de artigos, pelo que os *datasets* são geralmente maiores do que nos trabalhos com avaliação semi-automática e sobretudo manual. Verificamos, assim, num extremo, o trabalho com avaliação automática da qualidade da saúde em geral de Suzuki e Yoshikawa [61] com 484.146 artigos e, no outro extremo, os 10 artigos avaliados por Rajagopalan *et al.* [52]. Esta diferença pode ser atribuída ao facto dos trabalhos analisados na área da saúde serem efetuados por profissionais desta área, e portanto, sem as mesmas competências tecnológicas dos investigadores das áreas tecnológicas, origem da maior parte dos autores dos estudos da qualidade geral.

De modo a estimar um valor para a qualidade, os trabalhos de avaliação da qualidade em geral variam na sua abordagem, desde aqueles que se centram nas características do artigo em si, àqueles que avaliam o historial do artigo, nomeadamente as suas edições, e ainda aos que estudam a relação entre o autores e os artigos. Esta relação é também comum nos trabalhos de avaliação orientados para a saúde, onde é também muito utilizada a avaliação da inteligibilidade como forma de avaliar a qualidade da informação.

Existem, aparentemente, padrões em comum nos dois tipos de estudo. Um desses padrões é a utilização da pré-avaliação da Wikipédia de acordo com os respetivos WikiProjects para validação de resultados. São comumente utilizados os “Featured articles”, artigos pré-avaliados com avaliação máxima, como referência de artigos de elevada qualidade. É também comum serem descartados os artigos com classificação “Stub”, por se considerar que não reúnem condições para serem considerados artigos com informação passível de ser avaliada. Muitas vezes são utilizados os artigos classificados como “Start” quando se necessita de artigos avaliados com baixa qualidade em oposição aos “Featured Articles”. Outra semelhança entre os dois tipos de estudos reside nas versões utilizadas. A versão inglesa da Wikipédia é uma constante, ou como a única versão analisada ou como segundo idioma analisado. Este padrão, confirma a relevância da versão inglesa da Wikipédia em relação às restantes.

De modo a conseguir uma visão geral das medidas utilizadas nos trabalhos analisados, foi elaborada a Tabela 3.1. Nesta tabela constam os trabalhos acima descritos, divididos por área – geral ou de saúde – e as medidas utilizadas para avaliação da qualidade. Foram feitas algumas generalizações e adaptações no sentido de se obter uma classificação das medidas em três grupos diferentes. O primeiro grupo é o “Histórico” e agrega as medidas relacionadas com o histórico do artigo, desde a sua criação às suas posteriores edições e respetivos editores. O segundo grupo, denominado “Estrutural”, refere-se às medidas relacionadas com a estrutura do artigo, como as suas secções, referências, recursos multimédia ou *links*. Por fim, no grupo “Linguístico”, estão agregadas as medidas relacionadas com os aspetos linguísticos do artigo, sejam relacionadas com a inteligibilidade, seja por exemplo, com as palavras, frases e sílabas que o compõem.

Trabalhos Medidas		Geral									Saúde							Total de artigos
		Stvilla (2005)	Blumenstock (2008)	Wu (2010)	Xu (2011)	Suzuki (2013)	Li (2015)	De La Robertie (2015)	Marrese (2019)	Khan (2019)	Rajagopalan (2011)	Thomas (2013)	Conti (2014)	Modiri (2018)	Simpson (2018)	Swannakhan (2019)	Domingues (2019)	
Histórico	Idade do artigo																	6
	Editores únicos																	5
	Editores repetidos																	1
	Idade média dos editores																	2
	Upload médio dos editores																	2
	Número de editores anónimos																	1
	Edições																	9
	Edições por util. registados																	2
	Edições por util. anónimos																	3
	Edições por administradores																	2
	Edições menores																	2
	Tempo de reversão																	4
	Reversões de edições																	2
	Atualidade																	2
																		43
Estrutural	Conetividade																	2
	Links externos																	4
	Links internos quebrados																	2
	Links internos																	4
	Diversidade																	2
	Recursos multimédia																	4
	Templates																	1
	Infoboxes																	1
	Contagem de catálogos																	1
	Leituras futuras																	2
	Quantidade de secções																	2
	Quantidade de referências																	5
																		30
Linguístico	Quantidade de palavras																	8
	Rácio de palavras por frase																	2
	Variedade de palavras																	1
	Variedade de conceitos																	1
	Escalas de inteligibilidade																	8
	InfoNoise																	2
	Frases passivas																	1
	Sílabas																	2
	Erros ortográficos																	2
	Quantidade de frases																	1
																		28
Total de medidas		18	1	20	3	2	2	2	1	3	3	2	18	1	1	3	21	
		52									49							

Tabela 3.1: Medidas utilizadas nos trabalhos analisados

O trabalho de London *et al.* [35] não está incluído na tabela por ter sido feita uma avaliação totalmente manual sem nenhuma métrica objetivamente categorizável e comparável com as restantes.

Na tabela, é possível observar que o grupo mais frequente é o “Histórico” – com 43 medidas, seguido de “Estrutural” – com 30 medidas e por último “Linguístico” – com 28 medidas. De referir, ainda, que os trabalhos relacionados com a saúde utilizam em média maior número de medidas – 6,7, que os trabalhos de carácter geral – média de 5,8.

Da tabela, foram, ainda extraídos os conjuntos de medidas mais frequentes, que são maioritariamente referentes às categorias “Histórico” e “Estrutural”. A primeira medida é a idade do artigo. Esta é frequentemente utilizada quer nos trabalhos de avaliação geral, quer nos trabalhos de avaliação específica na saúde. Refere-se ao tempo decorrido desde a criação do artigo e é esperado que um artigo com mais idade esteja mais desenvolvido, mais maduro, por já ter sido potencialmente sujeito a maior revisão.

A segunda medida são os editores. O número de editores de um artigo é também comumente utilizado por ambos os tipos de trabalhos, referindo-se à quantidade de pessoas que fez edições ao artigo, esperando-se que um artigo mais popular, que atrai mais atenção ganhe mais qualidade ao longo do tempo. Existem trabalhos que distinguem os editores únicos dos editores que se repetem nas edições aos artigos.

A medida seguinte são as edições. A quantidade de vezes que um artigo é editado é também uma métrica frequente nos dois tipos de trabalhos e está de certo modo relacionado com os editores e a idade do artigo. Espera-se que um artigo que tenha sofrido mais edições ao longo do tempo se encontre num estado de desenvolvimento superior, quer por adição de conteúdo tornando-o mais completo, quer por correção de erros e adição de melhoramentos.

Em seguida, surgem as reversões, uma métrica mais utilizada nos trabalhos de âmbito geral, onde é dada importância sobretudo ao tempo que vai desde uma edição até à reversão da mesma. Uma reversão rápida significará à partida que a edição foi incorrecta, originando a sua correção tão rapidamente quanto maior a gravidade do erro e a facilidade da sua deteção. Está intimamente ligada ao vandalismo.

Outra medida são os *links*. Esta é uma métrica recorrente nos dois tipos de trabalhos, onde é feita distinção entre *links* internos, externos e se estes estão funcionantes ou não. Sendo os *links* uma das características da Wikipédia, esta métrica poderá indicar a qualidade de um artigo na medida em que um artigo com ligação a outros artigos ou outras páginas consegue fornecer informação complementar aos utilizadores.

A próxima medida é o tamanho do artigo. Esta é uma das medidas mais frequentes em ambos os tipos de trabalhos. Sendo uma medida fácil de calcular, parte do princípio que existe uma relação direta entre a quantidade de informação e a qualidade final desta. Existem trabalhos que fazem esta avaliação termos do número de frases, variedade de palavras, ou em rácios de palavras em frases, mas o mais frequente é utilizar a quantidade de palavras para quantificar o tamanho do artigo.

Outra medida frequente é a inteligibilidade. A inteligibilidade é utilizada pelos dois tipos de trabalhos, enquanto indicador frequente da qualidade dos artigos, embora seja mais frequente naqueles orientados para a saúde. Ainda que sejam utilizadas outras escalas, as duas mais frequentes são a Flesch Reading Ease e a Flesch-Kincaid Grade Level, destinadas a avaliar a dificuldade de compreensão durante a leitura e criadas em específico para a língua inglesa.

Em seguida, temos os recursos multimédia, uma medida valorizada em ambos os tipos de trabalho. Esta, refere-se à quantidade de recursos multimédia presentes, partindo do princípio que isto enriquece um artigo. Inclui imagens, vídeos e áudio, mas na maior parte das vezes apenas são contabilizadas as imagens.

Por fim, surgem as referências. A análise da quantidade de referências é particularmente recorrente nos trabalhos direcionados para a saúde, havendo inclusivamente alguns que avaliam não só a quantidade mas também a qualidade destas. É de esperar que mais e melhores referências sejam um indicador de uma qualidade superior.

Analisando os trabalhos já efetuados, suas diferenças, semelhanças e resultados, reúnem-se conhecimentos que permitirão definir as medidas e métricas a utilizar nesta dissertação de modo a conseguir avaliar de forma automática e com a maior precisão possível a qualidade da informação presente nos artigos da Wikipédia, com particular interesse na área da saúde.

Capítulo 4

Qualidade dos conteúdos multi-idioma de saúde na Wikipédia

Este capítulo abordará, em primeiro lugar, o propósito desta dissertação, isto é, da caracterização da Wikipédia como fonte de informação, com especial interesse na área da saúde e medicina, e nas eventuais diferenças que existam a nível da qualidade da informação em versões de diferentes idiomas. Serão, posteriormente, abordados os métodos utilizados para alcançar o propósito do problema, levantado nas vertentes previamente descritas.

4.1 Formulação do problema

O problema pode ser formulado em dois sub-problemas, o primeiro relativo à comparação da qualidade em diferentes idiomas, e o segundo relativo à avaliação automática da qualidade. Estes, são descritos nas secções seguintes.

4.1.1 Comparação da qualidade da Wikipédia em diferentes idiomas

Estando disponível em mais de 300 idiomas ativos, com mais a serem adicionados ao longo do tempo, e com a heterogeneidade já abordada na Secção 2.5, é importante avaliar se a qualidade da informação disponível para os utilizadores falantes dos diferentes idiomas é semelhante entre as diferentes versões ou se há disparidades, e se assim for, poder quantificá-la. Perante diferenças existentes no acesso a informação de saúde entre os falantes das diversas idiomas [3], a Wikipédia tem, assim, o potencial para reduzir ou acentuar este desequilíbrio. Este potencial traduz-se na informação disponibilizada ao público leigo em busca de informação, mas também aos próprios profissionais de saúde. Os profissionais com problemas no acesso a outros recursos de informação poderiam utilizar a Wikipédia como uma ferramenta útil no seu trabalho, colmatando os défices existentes. Torna-se, portanto, eminente, analisar o contributo da Wikipédia para a democratização do acesso à informação, em particular na área da saúde, um objetivo estratégico definido pela Organização das Nações Unidas, no seu mais recente plano de ações previstas até 2030 [8]. Neste contexto, esta dissertação procura responder a uma primeira pergunta:

Existem diferenças significativas na qualidade da informação em saúde, disponível na Wikipédia em diferentes idiomas?

4.1.2 Avaliação automática da qualidade

Enquanto importante fonte de informação, quer pela sua facilidade de acesso e pesquisa, quer pela sua abrangência de conteúdos e espírito colaborativo, a Wikipédia levantou, no entanto, desde sempre, questões sobre a fiabilidade e a própria qualidade da sua informação. Dada a enorme quantidade de conteúdos presentes na mesma, torna-se difícil avaliar manualmente a qualidade da informação oferecida, pelo que somos impelidos a criar métodos de avaliação automáticos, que tornem este processo eficiente.

Pelas consequências potencialmente nefastas, assume especial relevância avaliar a qualidade da informação na área médica, que conta já com mais de 223 mil artigos em 281 idiomas diferentes em Fevereiro de 2019 [94]. Com mais de 2 biliões de visualizações de artigos sobre saúde no ano de 2019 [92] demonstra-se o interesse do público pela Wikipédia enquanto fonte de informação, para os quais será de todo o interesse o conhecimento da fiabilidade da informação que obtêm na Wikipédia. O carácter semi-estruturado da Wikipedia e estruturado da Wikidata e também o facto de ser multilingue e abrangente em termos de assuntos cobertos, leva a que seja um recurso muito apetecível para aplicações informáticas que precisam desta informação, aumentando, ainda mais, a necessidade de se apurar se o uso desta fonte de informação se traduz numa transmissão de informação de qualidade. Esta avaliação pode dar, ainda, a quem colabora na sua edição uma visão do estado qualitativo da informação disponível, mostrando o caminho a seguir no futuro para uma constante melhoria da Wikipédia. No seguimento do exposto, surge uma segunda questão de investigação:

Pode a informação em saúde disponível na Wikipédia ser avaliada de forma mais específica?

4.2 Solução

Para responder ao problema levantado, surgiram caminhos diferentes. Poderia, por exemplo, ter sido seguida uma abordagem virada para a emergente inteligência artificial, como a utilização de *machine learning*. A abordagem seguida foi, no entanto, mais tradicional, tirando proveito do carácter semi-estruturado da Wikipédia. Nesse sentido, foi decidido seguir uma estratégia de exploração das métricas propostas por Stvilia *et al.* [58] abaixo descritas. Consideram-se como métricas, os instrumentos de avaliação de alto nível, compostas, por sua vez, por medidas – instrumentos de nível mais baixo, usados para descrever dimensões da informação.

4.2.1 Métricas de avaliação da qualidade

Stvilia *et al.* [58] definem sete métricas: *authority*, *completeness*, *complexity*, *informativeness*, *consistency*, *volatility* e *currency*. Estas métricas recorrem a 19 medidas, extraídas dos artigos

da Wikipédia e do respetivo histórico. Estas métricas foram utilizadas em ambas as fases e são descritas em seguida, tendo em conta as particularidades de cada fase.

4.2.1.1 Authority

Os autores definem a *authority* como o “grau de reputação de um objeto de informação numa determinada comunidade ou cultura” [58]. A fórmula para o seu cálculo é:

$$\text{Authority} = 0,2 * N.^{\circ} \text{ de editores únicos} + 0,2 * N.^{\circ} \text{ de edições} + 0,1 * \text{Conetividade} + 0,3 * N.^{\circ} \text{ de reversões} + 0,2 * N.^{\circ} \text{ de links externos} + 0,1 * N.^{\circ} \text{ de edições por utilizadores registados} + 0,2 * N.^{\circ} \text{ de edições por utilizadores anónimos}$$

O número de editores únicos corresponde aos autores distintos de edições ao artigo. A conectividade corresponde ao número de artigos ligados a um artigo em particular através de editores em comum. As reversões, correspondem ao número de reversões feitas a edições anteriores aos artigos. As edições podem ser feitas por utilizadores registado, utilizadores ou anónimos. Os *links* externos referem-se a ligações presentes ao longo do artigo que remetem para conteúdo externo à Wikipédia.

4.2.1.2 Completeness

A *completeness* é definida como “a granularidade do modelo do objeto de informação de acordo com uma ontologia IS-A de uso geral como a WordNet” [58]. A sua fórmula é a seguinte:

$$\text{Completeness} = 0,4 * N.^{\circ} \text{ de links internos quebrados} + 0,4 * N.^{\circ} \text{ de links internos} + 0,2 * \text{Comprimento do artigo}$$

Os *links* internos estão presentes no texto dos artigos e remetem para outras páginas dentro da Wikipédia. Os links quebrados, são aqueles que remetem para páginas atualmente indisponíveis. O comprimento do artigo reflete o tamanho do texto em caracteres.

4.2.1.3 Complexity

A *complexity* é definida como “o grau da complexidade cognitiva de um objeto de informação relativo a uma atividade particular”. A sua fórmula é:

$$\text{Complexity} = 0,5 * \text{Flesch reading ease} - 0,5 * \text{Kincaid grade level}$$

São utilizados dois índices de inteligibilidade para avaliar a complexidade. O primeiro é o Flesch reading ease e o segundo é o Kincaid grade level, projetados para a língua inglesa. Rudolf Flesch propôs, em 1948, a Reading ease readability formula [16]. O resultado da fórmula varia de 0 – texto extremamente difícil de ler, até 100 – texto muito fácil de ler. Posteriormente, em 1975, e a partir desta, Peter Kincaid desenvolveu, para a Marinha dos EUA, o Kincaid grade level

[28], com vista a avaliar a inteligibilidade dos textos, ao mesmo tempo que dá uma indicação da escolaridade necessária para a compreensão dos mesmos. O resultado desta é um valor correspondente a um nível de escolaridade dos Estados Unidos da América. Os resultados dos dois índices estão, assim, inversamente relacionados – um texto com uma pontuação baixa no Flesch reading ease corresponderá a uma pontuação alta no Kincaid grade Level. Ambas as fórmulas recorrem ao número de frases, palavras e sílabas no texto, e a relação entre elas para o seu cálculo final. Sendo que estes índices foram desenvolvidos para o inglês, para os restantes idiomas, existem adaptações dos mesmos, mas estas são escassas. Nesse sentido, para os outros idiomas além do inglês, foi decidido utilizar apenas o Flesch reading ease. Esta solução para o cálculo da *complexity* foi utilizada numa versão prévia de Stvilia *et al.* [59].

4.2.1.4 Informativeness

A definição de *informativeness* está ligada à “quantidade de informação que um objeto de informação contém” [58]. A sua fórmula é:

$$\text{Informativeness} = 0,6 * \text{InfoNoise} - 0,6 * \text{Diversidade} + 0,3 * \text{número de imagens}$$

O infoNoise baseia-se num trabalho anterior [100] e refere-se ao rácio entre a quantidade da informação depois de aplicar *stemming* e *stopping* e o tamanho do artigo antes de ser processado.

A diversidade corresponde ao rácio entre o número de editores únicos e o número de edições totais de um artigo. As imagens são um dos recursos multimédia disponíveis na Wikipédia.

4.2.1.5 Consistency

A *consistency* é definida como uma “medida da consistência da representação dos objetos de informação no que toca à sua estrutura, formato e precisão, de acordo com os padrões estabelecidos” [58]. A sua fórmula é:

$$\text{Consistency} = 0,6 * \text{Rácio de edições por administradores} + 0,5 * \text{Idade do artigo}$$

O rácio de edições por administradores corresponde à percentagem de edições feitas por administradores de entre o total de edições. A idade do artigo é avaliada, desde a sua criação, em dias.

4.2.1.6 Volatility

A *volatility* é definida, pelos seus autores, como “o tempo que a informação permanece válida” [58]. A sua fórmula é:

Volatility = Tempo médio de reversão das edições

Esta métrica reflete a mediana dos tempos de reversão das diversas edições aos artigos.

4.2.1.7 Currency

A *currency* é definida, pelos seus autores, como o “tempo decorrido desde a última revisão do artigo”. A sua fórmula é, simplesmente:

Currency = Atualidade

A atualidade corresponde à diferença entre a data de recolha dos dados e a data da última edição feita ao artigo, com o resultado final em dias.

4.3 Metodologia

O trabalho efetuado divide-se em duas grandes fases. A fase I corresponde à avaliação da qualidade dos artigos da Wikipédia relacionados com a saúde em diferentes idiomas. A fase II corresponde à proposta de medidas e consequentes métricas específicas de avaliação da qualidade de artigos da Wikipédia relacionados com saúde. A Figura 4.1 esquematiza de forma simplificada a metodologia utilizada, nas duas fases. As setas correspondem ao fluxo de informação.

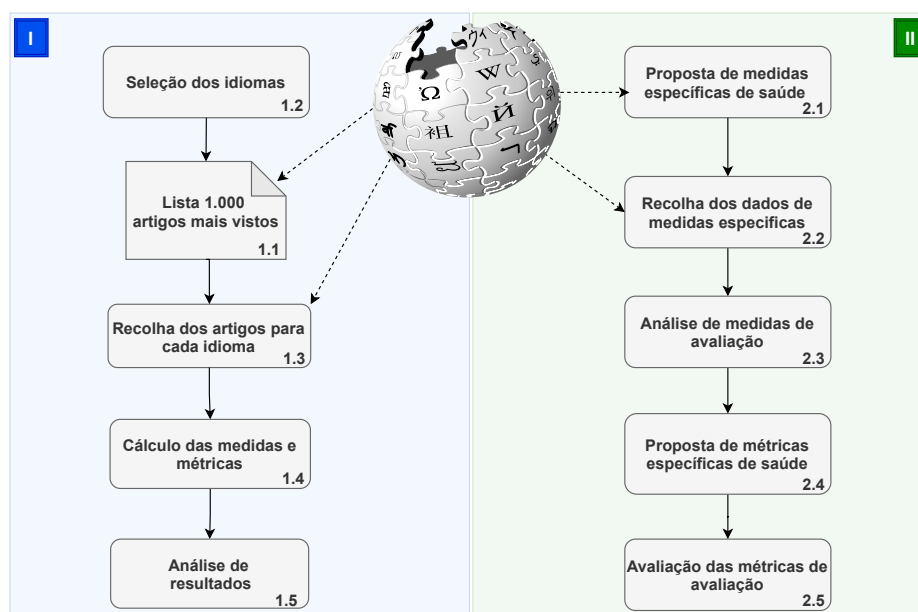


Figura 4.1: Metodologia utilizada

Relativamente à fase I, primeiro foi feita uma seleção dos idiomas a incluir, processo descrito na Secção 4.4.1. Em seguida, foi recolhida a lista de artigos relacionados com saúde. A etapa seguinte corresponde à recolha dos artigos da Wikipédia para cada idioma. Estas duas etapas estão descritas na Secção 4.4.2. Em seguida foram calculadas as medidas e métricas, descrito na Secção 5.1. Posteriormente, foi feita uma análise estatística dos resultados e sua discussão, presentes na Secção 5.2.

Na fase II, começou-se pela exploração de medidas específicas nos artigos da Wikipédia relacionados com saúde – transformadas em medidas, descrito na Secção 6.1. Em seguida, recolheram-se os dados necessários para o cálculo das medidas específicas; esta etapa está descrita na Secção 4.4.2. Na etapa seguinte, analisaram-se as medidas de avaliação – genéricas e as específicas de saúde, previamente propostas. Esta etapa está descrita na Secção 6.3. Partindo dos resultados desta etapa, foram propostas métricas específicas de saúde, descritas na Secção 6.3. Por fim, analisaram-se as métricas de avaliação específicas para saúde, previamente propostas, de acordo com a sua capacidade para avaliar a qualidade de artigos da Wikipédia relacionados com a saúde; os resultados desta análise são descritos na Secção 6.4.

4.4 Elaboração do *dataset*

A construção do *dataset* começou com a seleção dos idiomas que fazem parte do mesmo. Posteriormente foram recolhidos os dados dos artigos para os diferentes idiomas, necessários para o cálculo das métricas genéricas, na fase I. Houve ainda, na fase II, uma posterior recolha de dados necessários para a aplicação de métricas específicas de saúde, em inglês. Estas etapas são descritas em seguida.

4.4.1 Seleção dos idiomas

Foram selecionados idiomas disponíveis em versões da Wikipédia com pelo menos 100 milhões de falantes, enquanto idioma materno ou segundo idioma. Foram ainda selecionados sete outros idiomas pela sua importância histórica e cultural, ou médica, desde os tempos antigos: grego, italiano, coreano, turco, persa, hebreu e catalão. Na Tabela 4.1 podem-se observar o número de falantes para cada idioma do *dataset*, ordenados por ordem decrescente do número total de falantes.

O inglês é o idioma com mais falantes, sendo esta diferença mais acentuada enquanto segundo idioma. Para a fase II, e tendo em conta os objetivos, foi apenas selecionado o inglês. O chinês surge em segundo lugar, a nível de número de falantes totais, tendo mesmo o maior número de falantes enquanto idioma materno. Em terceiro lugar, surge o hindi, mas já com uma grande diferença no número total de falantes em relação aos dois primeiros. O catalão e o hebreu ocupam o dois últimos lugares, com apenas 9 milhões de falantes.

Além da heterogeneidade entre o número de falantes, existem ainda acentuadas diferenças linguísticas entre os idiomas selecionados. Os idiomas selecionados estão classificados em sete diferentes famílias linguísticas [12]. A maior parte dos idiomas faz parte da família indo-europeia:

Tabela 4.1: Número de falantes de cada idioma do *dataset*

	Idioma materno	Segundo idioma	Total
Inglês (en)	369,9 milhões	978,2 milhões	1,348 milhões
Chinês (zh)	921,2 milhões	198,7 milhões	1,120 milhões
Hindi (hi)	342,2 milhões	258,3 milhões	600 milhões
Árabe (ar)	–	–	274 milhões
Bengali (bn)	228,7 milhões	39,0 milhões	268 milhões
Francês (fr)	79,6 milhões	187,4 milhões	267 milhões
Russo (ru)	153,7 milhões	104,3 milhões	258 milhões
Português (pt)	232,4 milhões	25,2 milhões	258 milhões
Urdu (ur)	69,0 milhões	161,0 milhões	230 milhões
Indonésio (id)	43,6 milhões	155,4 milhões	199 milhões
Alemão (de)	76,6 milhões	58,5 milhões	135 milhões
Japonês (ja)	126,3 milhões	121,500	126 milhões
Turco* (tr)	82,2 milhões	5,9 milhões	88 milhões
Persa* (fa)	56,3 milhões	17,9 milhões	74 milhões
Coreano* (ko)	–	–	82 milhões
Italiano* (it)	64,8 milhões	3,1 milhões	68 milhões
Grego* (el)	–	–	13 milhões
Catalão* (ca)	–	–	9 milhões
Hebreu* (he)	–	–	9 milhões

* idiomas selecionados pela sua importância cultural ou médica

bengali, catalão, alemão, grego, inglês, persa, francês, hindi, italiano, português, russo e urdu. O árabe e o hebreu pertencem à família afro-asiática. O indonésio pertence à família austronésia, o japonês à japônica, o coreano à coreânica, o turco à turcomana e o chinês à sino-tibetana.

Outra grande distinção linguística reside nos sistemas de escrita. Entre os 19 idiomas, existem 11 sistemas de escrita diferentes. O alfabeto latino, ou romano, é partilhado pelo inglês, catalão, francês, português, indonésio, alemão, e italiano. Este alfabeto possui 21 letras de base, mas absorve variações de diferentes idiomas, como por exemplo do alemão, que possui 4 letras extra. O turco, utiliza o alfabeto turco, que é uma variante do alfabeto latino, possuindo 29 letras. O árabe, o persa, o urdu, e o hebreu, adotam um sistema de escrita abjad. Apesar de se poder denominar de alfabeto, todas as suas letras são consoantes. O árabe conta com 28 letras, o persa com 32 e o urdu e o hebreu, contam com 40 e 22, respetivamente. As vogais podem ser formadas por diacríticos colocados sobre, ou sob, as letras e não existe o conceito de maiúsculas e minúsculas. O bengali adota o alfabeto bengali. Tal como o devanágari, adotado pelo hindi, são abugidas, ou alfassilabários. O bengali conta com 51 letras, e o devanágari conta com 48. O russo utiliza uma variante do alfabeto cirílico, composta por 33 letras – 21 consoantes, 10 vogais e dois sinais. O coreano utiliza o Hangul, um alfabeto com 14 consoantes e 10 vogais. As letras são escritas em blocos e não sequencialmente; cada um destes blocos silábicos contém duas a seis letras. O grego, utiliza o alfabeto grego, derivado do alfabeto fenício. Foi o primeiro sistema de escrita a utilizar letras distintas para as vogais e as consoantes. O chinês e o japonês não possuem um alfabeto propriamente dito. No chinês, os caracteres – Hanzi são logogramas – cada carater tem um

conceito associado. As palavras são formadas por associação de caracteres e respetivos conceitos. O japonês é baseado em três sistemas de escrita, o hiragana, o katakana e o kanji – derivado do chinês. Os sistema de escrita árabe, hebreu, persa e urdu, ao contrário dos restantes, usam a escrita da direita para a esquerda.

4.4.2 Recolha dos artigos multi-idioma

Os artigos a recolher foram aqueles presentes numa lista criada e mantida pelo WikiProject Medicine [93]. Esta lista contém os 1.000 artigos de saúde mais vistos na Wikipédia inglesa. Sendo uma lista constantemente atualizada, garante-se assim que inclua artigos recentes e relevantes, como por exemplo sobre a pandemia de COVID-19.

Numa primeira fase da recolha, foi recolhida uma amostra de artigos da lista para cada idioma. Esta fase serviu para validar os dados recolhidos, permitindo também uma afinação dos métodos de recolha. Primeiro, foram recolhidos os artigos em inglês da referida lista. Em seguida, foram recolhidos os dados necessários para os outros idiomas, de forma iterativa.

Foi utilizada a API da MediaWiki para recolher os dados e metadados do artigo, seguindo a abordagem de Domingues e Teixeira Lopes [9]. Foi através desta que foi recolhido o histórico de revisões, incluindo as edições e os seus autores; foram ainda recolhidos os *links* dos idiomas, *links* internos e *links* externos. A quantidade de imagens foi recolhida a partir da linguagem *markup* do artigo. Para calcular medidas relacionadas com o texto, como o índice de legibilidade Flesch Reading Ease, o infoNoise ou o comprimento do artigo, é necessário obter texto pleno. Nesse sentido, removeram-se todas as *markups* do conteúdo do artigo para obter um ficheiro de texto pleno e este foi usado no cálculo das mesmas.

Surgiram desafios na recolha de dados devidos à uma heterogeneidade entre os idiomas escolhidos, acima exposta. Também há heterogeneidade entre as diferentes versões da Wikipedia para cada idioma. Além disso, há também heterogeneidade dentro de cada versão da Wikipédia, pois as edições são feitas por vários utilizadores, que nem sempre cumprem os padrões estabelecidos, quando existentes.

4.4.3 Recolha de dados para as métricas específicas

Na fase II, foi necessário recolher informação adicional para o cálculo das medidas e respetivas métricas específicas de saúde, apenas em inglês.

Para recolher apenas *templates* e *infoboxes* relacionados com saúde, foi gerada uma lista para cada um deles, a partir das categorias e subcategorias da lista de *templates* médicos da Wikipedia¹ e da lista de *infoboxes* médicos². Foi recolhida a lista de administradores, ativos e inativos do WikiProject Medicine, a partir da respetiva página na Wikipédia³. A partir da página da Healthcare

¹https://en.wikipedia.org/wiki/Category:Medicine_templates

²https://en.wikipedia.org/wiki/Category:Medicine_infobox_templates

³https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Participants

Translation Task Force foi recolhida a lista de artigos traduzidos⁴. Os códigos médicos foram recolhidos e contabilizados para cada artigo. A Wikipedia reúne uma lista de códigos que podem ser incluídos nos *templates* de artigos⁵. Foi ainda necessário obter uma lista de *links* reputados; para isso, foi feito *scraping* dos *links* sugeridos pelo National Institute of Health⁶, parte do U.S. Department of Health and Human Services. Recolheu-se ainda, uma lista de secções recomendadas pelo WikiProject Medicine⁷. Um dos dados necessários foi o nível de qualidade dos artigos definido pelo WikiProject Medicine, de acordo com a Tabela 2.1. Para obter este dado, foi recolhida a lista de artigos de cada nível de classificação relevante⁸.

Os dados recolhidos foram guardados numa base de dados relacional, representada na Figura 4.2.

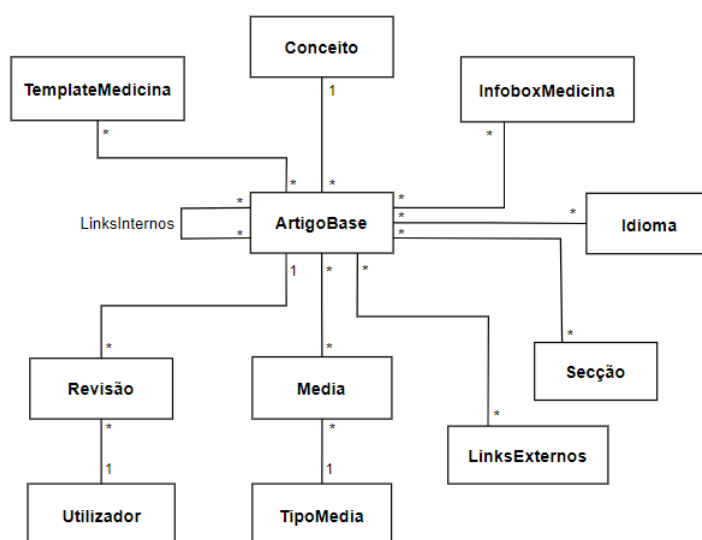


Figura 4.2: Estrutura da base de dados. Adaptado de: Domingues e Teixeira Lopes [9]

4.4.4 Caraterização do dataset

Como alguns artigos só estão disponíveis nalguns idiomas, o *dataset* final contém um total de 14.456 artigos. A distribuição dos artigos em função do idioma é mostrado na Figura 4.3. Esta figura inclui, também, a distribuição do número de artigos para cada versão da Wikipédia, para comparação. A Figura 4.3 mostra que o inglês é o único idioma com 1.000 artigos, o que significa que nenhum outro idioma possui a versão correspondente para todos os artigos da lista. Também são evidentes as diferenças entre os idiomas, onde alguns têm apenas cerca de metade, ou até menos, do número total de artigos, como o urdu. Analisando a relação entre o número de artigos do

⁴https://en.m.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Translation_task_force

⁵https://en.wikipedia.org/wiki/Template:Medical_resources

⁶<https://www.nlm.nih.gov>

⁷https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Medicine-related_articles

⁸https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Assessment#Quality_assessment

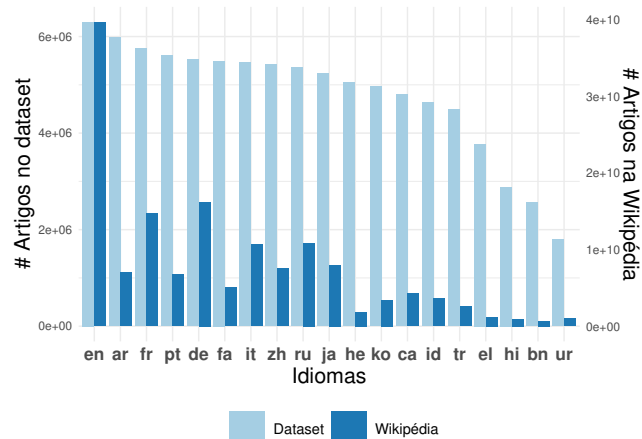


Figura 4.3: Número de artigos por idioma

conjunto de dados e o número total de artigos em cada versão da Wikipédia, podemos observar diferenças nas distribuições. Dado que o conjunto de dados contém apenas artigos relacionados com a saúde, esses dados sugerem que o tamanho de cada versão da Wikipédia não está diretamente relacionado com o número de artigos relacionados à saúde. Esses dados também podem apontar diferenças na importância dada à área da saúde em cada versão da Wikipédia.

A distribuição dos artigos, em inglês, pelos níveis de qualidade, já referidos, é a seguinte: FA (2.9%), GA (8.2%), B (42.8%), C (35%), Start (8.6%), Stub (0.4%), e List (1.4%). Os trabalhos existentes [58, 59] normalmente descartam os artigos classificados como *stub* e aqui foi usada a mesma abordagem. Também foram descartados os artigos classificados como *list*, correspondentes a 1,4% dos artigos do *dataset* em inglês, pela sua natureza que não é coerente com os restantes artigos.

Para além de classificar os artigos por qualidade, o WikiProject Medicine, classifica-os também por importância. É, assim, avaliada a prioridade do artigo para cada WikiProject. No WikiProject Medicine, “o objectivo da classificação por importância é dirigir os esforços de melhoria do artigo do projecto para os artigos mais importantes, e incidentalmente fornecer uma conveniente lista restrita de tópicos importantes para os leitores que estão interessados em medicina em geral”⁹. No *dataset* em inglês, a importância dos artigos é classificada como: 7.2% *Top importance*; 25.4% *High importance*; 43.3% *Mid importance*; 23.4% *Low importance*; 0.7% dos artigos não foram classificados quanto à sua importância.

⁹https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Assessment

Capítulo 5

Comparação multi-idioma da qualidade de artigos de saúde da Wikipédia

Neste capítulo será descrita a fase do trabalho relativa à comparação da qualidade dos artigos de saúde na Wikipédia nos vários idiomas. Primeiro são abordados os cálculos das métricas e respetivas medidas. Em seguida é feita uma análise estatística, onde se descreve a estratégia usada, são expostos os resultados e por fim discutem-se os resultados. O capítulo termina com uma conclusão relativa a esta fase.

5.1 Cálculo de medidas e métricas

O cálculo de cada métrica implica um cálculo prévio de cada uma das medidas que a compõem. A Tabela 5.1 compila as métricas, medidas, a origem dos dados para o seu cálculo e as bibliotecas, nas medidas utilizadas.

Para o cálculo da métrica *authority*, foi necessário calcular o o número de edições, reversões, *links* e respetivos autores. Estes dados foram obtidos diretamente através do histórico dos artigos recolhidos. Existe, no entanto, algo a ter em conta ao contabilizar as reversões – a contabilização destas, depende da identificação da edição como sendo uma reversão, por parte do autor da mesma. Nem sempre os autores têm este e outros cuidados com a estrutura dos artigos, e das edições em particular. A conectividade é outra das medidas, e foi obtida extraindo os editores de cada artigo e os artigos editados por estes, recorrendo ao histórico; os dados são depois cruzados entre si. Esta métrica tem a limitação de poder ser apenas calculada com base nos artigos presentes na base de dados. Este facto foi apontado por Domingues e Teixeira Lopes [9], sendo também referido o facto desta limitação poder, inclusive, fortalecer o conceito de autoridade, uma vez que os artigos tidos em conta estão necessariamente relacionados com a área da saúde.

Para a *completeness*, são contabilizados os *links* a partir do *markup* dos artigos. Posteriormente, obtêm-se aqueles *links* que estão quebrados a partir de chamadas à API. O comprimento do artigo é contabilizado a partir do ficheiro de texto pleno, previamente criado.

Tabela 5.1: Processo de cálculo das medidas

Métrica	Medida	Origem dos dados	Bibliotecas
Authority	Editores únicos	Histórico do artigo	
	Edições		
	Reversões		
	Links externos		
	Edições por ut. registados		
	Edições por ut. Anónimos		
Completeness	Conetividade	Markup do artigo	
	Links internos		
	Links internos quebrados		
	Comprimento do artigo		
Complexity	Flesch Reading Ease	Texto pleno	Textstat
Informativeness	Infonoise	Markup do artigo	Spacy
		Texto pleno	Snowballstemmer
	Diversidade	Histórico do artigo	Bengla-stemmer
			PersianStemmer
			KoNLPy
Consistency	Imagens	Markup do artigo	
	Edições por administradores	Histórico do artigo	
	Idade do artigo	Lista de administradores	
Volatility	Tempo médio de reversões	Histórico do artigo	
Currency	Atualidade	Histórico do artigo	

Para a *complexity*, considerando apenas o índice Flesch reading ease, este foi obtido, também a partir de texto pleno. Para obter os elementos do texto necessários para o cálculo, foi utilizada uma biblioteca de *Python* – *Textstat* [20]. Para inglês, o francês, alemão e italiano, o valor é obtido diretamente a partir da biblioteca referida. Para o português, recorreu-se a uma adaptação, que consiste em adicionar 42 ao valor do Flesch reading ease, de acordo com o trabalho de Teresa Martins *et al.* [38]. Para o turco, foi utilizada a Fórmula de Atesman [10]. Para os restantes idiomas, não foi encontrada uma implementação com resultados satisfatórios.

A *informativeness* implica o cálculo do infoNoise. Para o calcular, é necessário aplicar *stemming* e *stopping* ao texto do artigo. Para estes processos, recorreu-se a bibliotecas de *Python*. Para a seleção das mesmas, foram pesquisadas as opções existentes e o facto de ser necessário analisar diferentes idiomas, mantendo a consistência ao lidar com as suas especificidades. Para o processo de *stopping* e *tokenization* foi utilizada a biblioteca *spaCy* [13], para todos os idiomas. Para o processo de *stemming*, foi utilizada a biblioteca *Snowballstemmer* [19] para os idiomas disponíveis: inglês, português, francês, alemão, italiano, espanhol, russo, árabe, indonésio, grego, turco e hindi. Para o bengali, foi utilizada a biblioteca *Bangla-stemmer* [17]. Para o persa, foi utilizada a biblioteca *PersianStemmer* [18] e para o coreano, foi utilizada a biblioteca *KoNLPy* [47]. Para o chinês, o japonês, o urdu e o hebreu não foi encontrada uma implementação de um *stemmer* funcional. Outra medida integrante da *informativeness* é a diversidade. Esta medida implica a obtenção dos editores únicos e do número de edições; o cálculo destes dados foi explicado acima, para a *authority*. O número de imagens, a última medida, foi contabilizado entre os diferentes objetos de média presentes no artigo, filtrando o tipo de ficheiro, através da extensão destes.

Para a *consistency*, é necessário calcular o número de edições feitas por administradores. Para

se obterem dados mais fidedignos, foram tidos em conta, não só os administradores ativos, mas todos os administradores de cada idioma. Esta medida avalia se uma edição é feita por um utilizador que é, ou já foi, administrador, não necessariamente à altura da edição. A idade do artigo corresponde à diferença entre a data de criação do artigo e a data de recolha dos dados.

A *volatility* implica apenas o tempo médio das reversões, em minutos. O cálculo das reversões e suas fragilidades já foram descritos acima, relativamente à métrica *authority*. Esta métrica tem, contudo, uma particularidade – quando não existem reversões num artigo, o valor da mediana vai ser 0, da mesma forma que também é 0 quando, existindo reversões, o tempo mediano destas seja, de facto, 0. Esta situação verifica-se nos trabalhos de Domingues e Teixeira Lopes [9] – para o português e de Stvilia *et al.* [58] – para o *dataset* de idiomas aleatórios.

Por fim, para a *currency*, é calculada a diferença entre a data de recolha dos dados e a data da última edição feita ao artigo, a partir do histórico.

5.2 Análise estatística

Calculadas as métricas, é feita uma análise dos estatística dos resultados. A estratégia para o conseguir e os resultados em si são expostos nas próximas secções.

5.2.1 Estratégia de análise

Os vários idiomas são comparados em termos de medidas e métricas. Como a maioria das métricas e medidas não seguem uma distribuição normal em cada idioma, é utilizada a mediana como medida de tendência central e o intervalo interquartil como medida de dispersão. Como não se verificaram as condições para análise de variância (ANOVA), aplicou-se o teste de Kruskal-Wallis por postos para perceber se havia diferenças significativas entre os idiomas em cada medida e cada métrica. Foram efetuados testes *post-hoc*, nomeadamente o teste de pares de Dunn, com valores-*p* ajustados pelo método de Holm, para identificar as diferenças significativas.

Para análise da métrica, foi elaborada uma tabela onde estão presentes os idiomas ordenados por ordem decrescente da mediana. Além da mediana de cada idioma, está também presente o intervalo interquartil (IIQ). Está ainda presente a lista de idiomas de qualidade inferior, com um nível de significância *alfa* entre 0,01 inclusive e 0,05 para os valores marcados com asterisco e um nível igual ou inferior a 0,01 para os restantes; no final da linha está contabilizado o total de idiomas significativamente inferiores. Está ainda representado o valor do teste de qui-quadrado (χ^2) e respetivo valor-*p*. Para análise das medidas, foram construídos diagramas de caixa. Para uma visualização conveniente, os *outliers* não são mostrados. Um “X” marca o valor da média em cada um deles. Foi ainda adicionada uma visualização de partes específicas do diagrama, quando estas são pouco reveladoras.

5.2.2 Resultados

São, em seguida, descritos os resultados para cada uma das métricas e respetivas medidas.

5.2.2.1 Authority

A Tabela 5.2 representa os dados da métrica *authority*. Da sua análise, observa-se que o inglês se destaca positivamente dos restantes idiomas; além de ser o idioma com maior mediana é também o idioma com maior dispersão de valores, traduzido no maior IIQ entre todos os idiomas. Em segundo e terceiro lugares temos, respetivamente, o alemão e o russo. Em último lugar surge o bengali, com uma mediana de apenas cerca de um terço daquela do inglês. O teste de Kruskal-Wallis indica que existem diferenças significativas entre os idiomas. Quando analisadas as diferenças significativas entre cada idioma, verifica-se, no entanto, que o terceiro classificado – russo, não é significativamente diferente do quarto classificado – italiano. Da mesma forma, também se verifica que existem três idiomas – urdu, indonésio e grego, que não são significativamente diferentes do último classificado – bengali.

Tabela 5.2: Avaliação da qualidade dos idiomas para a métrica *authority*

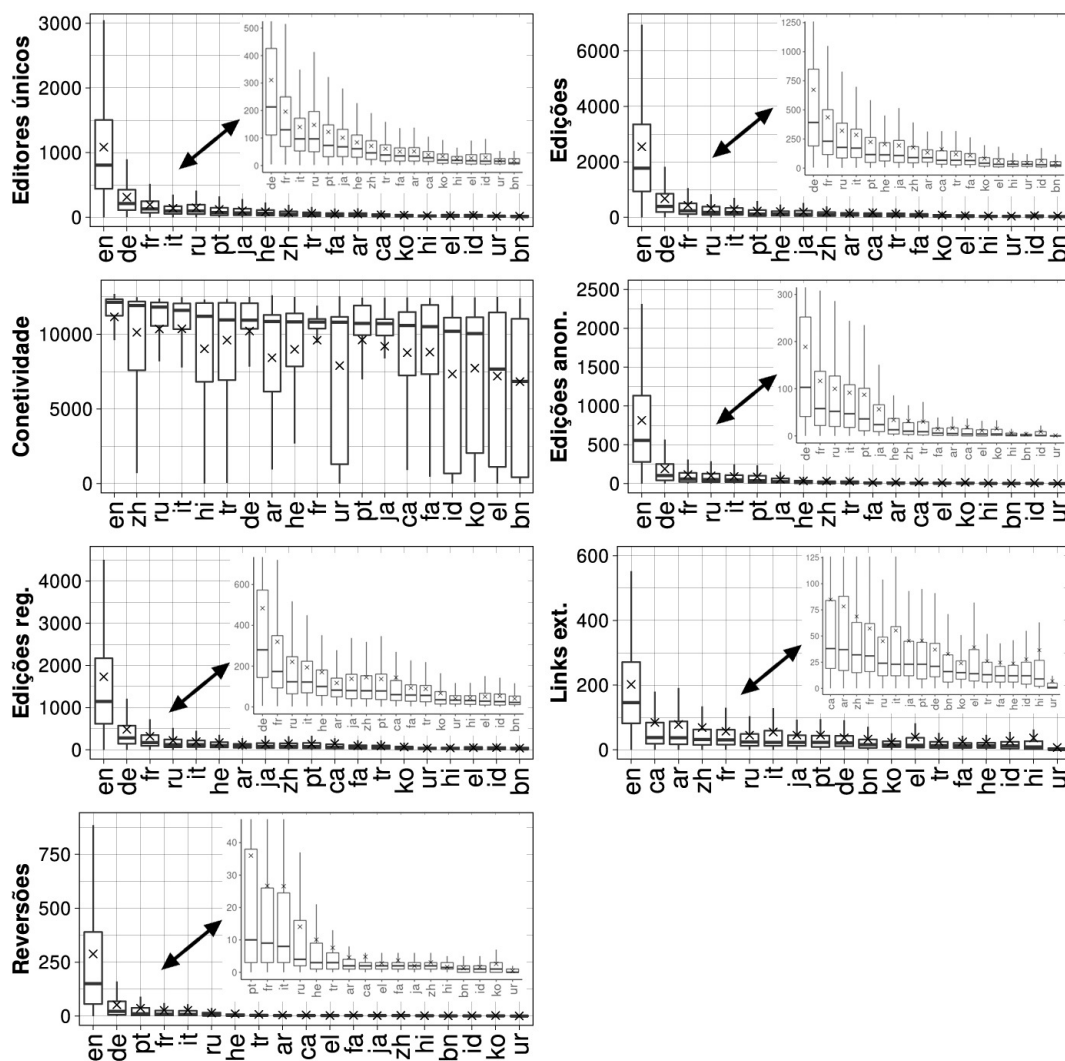
		Mediana	IIQ	Idiomas significativamente inferiores																	#	
Inglês	en	2033,1	1196,7	de	ru	it	zh	fr	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn	18
Alemão	de	1315,5	416,7	ru	it	zh	fr	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn		17
Russo	ru	1250,8	265,6	zh	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn					14
Italiano	it	1240,2	254,7	zh*	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn					14
Chinês	zh	1230,4	344,5	hi	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn							12
Francês	fr	1189,8	233,8	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn						13
Hindi	hi	1159,6	546,1	pt	ur	id	ko	el	bn													6
Português	pt	1152,6	273,5	ar	ja	ca	ur	fa	id	ko	el	bn										9
Turco	tr	1148,5	539,6	ur	fa	id	ko	el	bn													6
Hebreu	he	1139,3	413,5	ur	fa*	id	ko	el	bn													6
Árabe	ar	1130,3	603,0	ur	id	ko	el	bn														5
Japonês	ja	1128,4	186,8	ur	id	ko	el	bn														5
Catalão	ca	1101,6	468,0	ur	id	ko	el	bn														5
Urdu	ur	1096,6	994,0	fa																		1
Persa	fa	1087,2	490,4	id	ko	el	bn															4
Indonésio	id	1034,5	1086,8																			0
Coreano	ko	1024,3	956,2	bn																		1
Grego	el	800,8	1069,9																			0
Bengali	bn	710,1	1070,2																			0
χ²		3543,3																				
Valor-p		<2,2e-16																				

* nível de significância $0,01 < p \leq 0,05$, nível de significância $p \leq 0,01$ para os restantes valores, IIQ: Intervalo interquartil, #: total de idiomas significativamente inferiores

Nos diagramas de caixa das medidas, na Figura 5.1, pode-se observar que o inglês lidera em todas elas. O número de reversões varia consideravelmente entre os idiomas, com uma dominância clara do inglês. O inglês é ainda o idioma que apresenta o maior IIQ.

5.2.2.2 Completeness

Na Tabela 5.3 verifica-se que o inglês surge novamente como líder entre os idiomas para a métrica *completeness*. Em segundo lugar surge novamente o alemão e em terceiro lugar surge o francês. Os últimos lugares são ocupados pelo chinês, coreano e urdu, este último com uma mediana quase 20 vezes inferior à do inglês. O inglês volta a apresentar-se como o idioma com maior IIQ. O teste de Kruskal-Wallis revela novamente que existem diferenças significativas entre

Figura 5.1: Distribuição das medidas da métrica *authority*

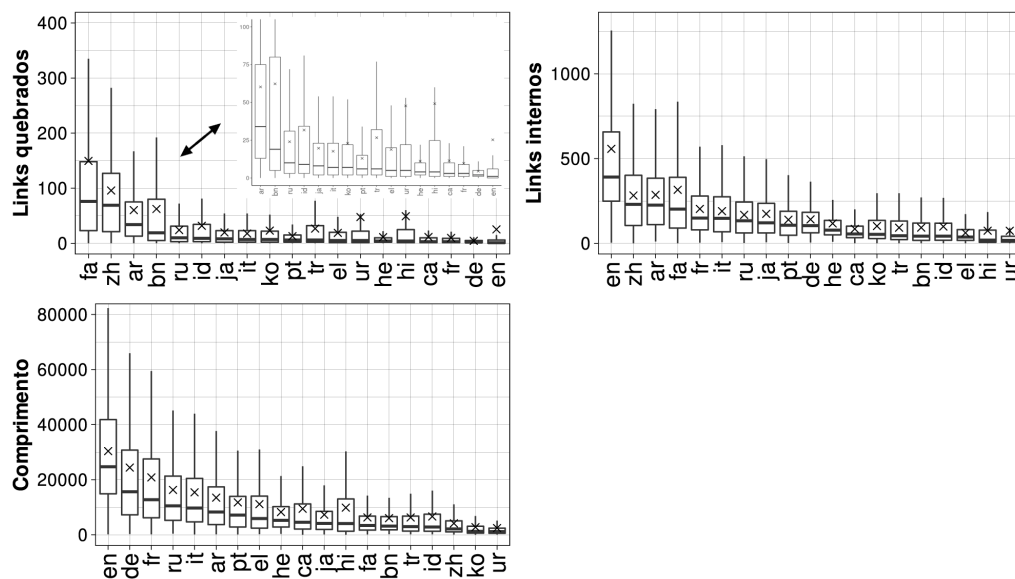
os idiomas, para a *completeness*. O francês não apresenta, contudo, diferenças significativas do russo, que surge em quarto lugar. A última posição é, desta vez, ocupada pelo urdu, não havendo diferenças significativas para o coreano, na penúltima posição.

Analisando a Figura 5.2, podemos observar que o inglês volta a ser o idioma dominante, não sendo o idioma com maior mediana apenas nos *links* internos quebrados. Nesta medida, o inglês situa-se mesmo na última posição, sendo que o topo pertence ao persa. Quando cruzamos os *links* internos quebrados com o número total de *links* internos, observa-se que geralmente, os idiomas com maior número de *links* internos têm também maior número de *links* quebrados. O inglês apresenta-se, no entanto, como uma exceção, pois apesar de ser o idioma com maior número de *links* internos, é o idioma com menos *links* quebrados. Relativamente ao comprimento do artigo, o inglês consegue uma mediana mais de 150% superior à do segundo classificado – o alemão, e quase 2.200% superior à do último classificado – o urdu.

Tabela 5.3: Avaliação da qualidade dos idiomas para a métrica *completeness*

		Mediana	IIQ	Idiomas significativamente inferiores																	#	
Inglês	en	5132,4	5420,3	de	fr	ru	it	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur	18
Alemão	de	3172,8	4705,0	fr*	ru	it	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur		17
Francês	fr	2619,3	4362,4	it	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur				15
Russo	ru	2154,0	3262,3	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur					14
Italiano	it	2042,0	3176,2	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur						13
Árabe	ar	1810,8	2799,0	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur						13
Português	pt	1471,4	2244,2	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur								11
Grego	el	1203,6	2319,7	ja	fa	hi	bn	id	tr	zh	ko	ur										9
Hebreu	he	1080,2	1511,2	ja*	fa	bn	id	tr	zh	ko	ur											8
Catalão	ca	936,0	1853,8	bn	id	tr	zh	ko	ur													6
Japonês	ja	903,2	1346,4	id	tr	zh	ko	ur														5
Persa	fa	868,4	1179,3	id	tr	zh	ko	ur														5
Hindi	hi	830,9	2350,6	id*	tr	zh	ko	ur														5
Bengali	bn	676,7	995,9	ko	ur																	2
Indonésio	id	629,8	1312,4	ko	ur																	2
Turco	tr	625,6	1083,5	ko	ur																	2
Chinês	zh	583,0	846,6	ko	ur																	2
Coreano	ko	301,2	534,7																			0
Urdu	ur	271,3	422,6																			0
χ2		4410,6																				
Valor-p		<2,2e-16																				

* nível de significância $0,01 < p \leq 0,05$, nível de significância $p \leq 0,01$ para os restantes valores. IIQ: Intervalo interquartil, #: total de idiomas significativamente inferiores

Figura 5.2: Distribuição das medidas da métrica *completeness*

5.2.2.3 Complexity

A Tabela 5.4 mostra apenas os valores dos seis idiomas para os quais foi possível calcular a *complexity*. Na tabela, o português surge em primeiro lugar, seguido do turco. O último e o penúltimo lugares são ocupados pelo italiano e inglês, respetivamente. O italiano consegue apenas 17% da mediana do português. Apesar de ser o primeiro classificado, o português não é o idioma com maior dispersão de valores, sendo o turco que alcança um maior IIQ. Os outliers desta

métrica, não mostrados, são maioritariamente relativos a artigos, em diversos idiomas, que são listas, como por exemplo, listas de mortos por COVID-19 resultando num resultado enviesado do índice Flesch reading ease. O teste Kruskal-Wallis revela diferenças significativas entre os idiomas para esta métrica. O primeiro classificado – português é, de facto, significativamente superior a todos os outros idiomas.

Tabela 5.4: Avaliação da qualidade dos idiomas para a métrica *complexity*

		Mediana	IIQ	Idiomas significativamente inferiores					#
Português	pt	76	18	tr	de	fr	en	it	5
Turco	tr	62	34	de	en	it			3
Alemão	de	56	10	fr	en	it			3
Francês	fr	45	18	en*	it				2
Inglês	en	34	14	it					1
Italiano	it	13	26						0
χ^2		11373							
Valor- <i>p</i>		<2,2e-16							

* nível de significância $0,01 < p \leq 0,05$, nível de significância $p \leq 0,01$ para os restantes valores, IIQ: Intervalo interquartil, #: total de idiomas significativamente inferiores

A única medida da *complexity* é o Flesch reading ease. A respetiva distribuição é mostrada na Figura 5.3. Nesta, podemos confirmar os dados da Tabela 5.4.

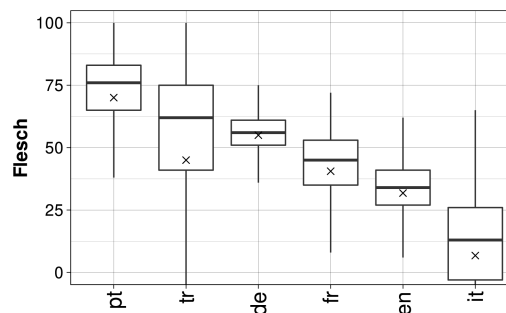


Figura 5.3: Distribuição da medida Flesch reading ease

5.2.2.4 Informativeness

A Tabela 5.5 mostra os resultados da *informativeness* para os idiomas onde foi possível calcular o infoNoise. Da sua análise, podemos concluir que o inglês se destaca novamente pela positiva, conseguindo mais do triplo da mediana do segundo classificado – árabe. O inglês é, também novamente, o idioma com maior IIQ. O hindi ocupa, de novo, o último lugar, de entre os idiomas considerados. O teste de Kruskal-Wallis revela diferenças significativas entre os idiomas para a informativeness. O inglês é significativamente superior a todos os outros idiomas. O indonésio, o persa, o turco e o hindi não são significativamente superiores a nenhum idioma.

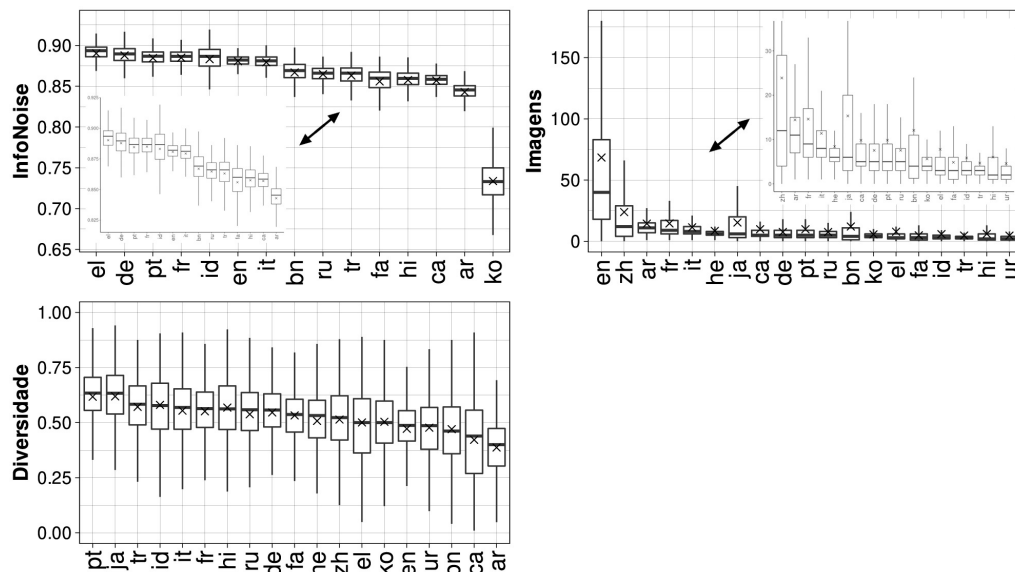
Analisando as medidas, de acordo com a Figura 5.4, o inglês ocupa o primeiro lugar apenas no número de imagens. Na diversidade, onde um valor inferior se traduz em maior qualidade, o árabe ocupa o primeiro lugar, o último lugar cabe ao português. O grego consegue a mediana mais alta

Tabela 5.5: Avaliação da qualidade dos idiomas para a métrica *informativeness*

		Mediana	IIQ	Idiomas significativamente inferiores														#
Inglês	en	12,38	19,48	ar	fr	it	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	14
Árabe	ar	3,53	2,40	fr	it	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi		13
Francês	fr	3,01	3,20	it	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi			12
Italiano	it	2,57	1,83	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi				11
Catalão	ca	1,83	1,78	de	bn	ko	el	id	fa	tr	hi							8
Russo	ru	1,68	1,34	ko	el	id	fa	tr	hi									6
Alemão	de	1,67	1,88	ko	el	id	fa	tr	hi									6
Português	pt	1,65	1,84	ko	el	id	fa	tr	hi									6
Bengali	bn	1,45	2,87	ko	id	fa	tr	hi										5
Coreano	ko	1,38	0,90	fa*	tr	hi												3
Grego	el	1,25	1,37	id	fa	tr	hi											4
Indonésio	id	1,12	1,14															0
Persa	fa	1,10	1,33															0
Turco	tr	1,07	0,81															0
Hindi	hi	0,82	1,61															0
χ^2		4002,1																
Valor-p		<2,2e-16																

* nível de significância $0,01 < p \leq 0,05$, nível de significância $p \leq 0,01$ para os restantes valores, IIQ: Intervalo interquartil, #: total de idiomas significativamente inferiores

para o infoNoise, onde o coreano fica em último lugar. Nota-se uma distribuição para o coreano distinta das restantes, o que pode estar associado ao facto de ter sido usado um *stemmer* diferente da maioria dos restantes idiomas.

Figura 5.4: Distribuição das medidas da métrica *informativeness*

Tendo em conta o facto do infoNoise condicionar a comparação entre os idiomas, deixando alguns de fora, foi calculado o infoNoise sem o *stemming*. Foram, portanto, aplicados apenas os processos de *stopping* e *tokenization*, e posteriormente calculada a *informativeness* com o valor resultante. Os resultados estão presentes na Tabela 5.6. O primeiro dado a destacar é o facto de apenas haver diferença no valor do IIQ para o inglês, em relação aos valores da Tabela 5.5, onde

foi aplicado *stemming*, para valores arredondadas às centésimas. Esta circunstância pode dever-se ao peso que o *stemming* tem no resultado do cálculo do infoNoise, associado ao peso do próprio infoNoise no resultado final da métrica. Relativamente à posição dos idiomas, verifica-se que o chinês passa a ocupar o segundo lugar, o hebreu e o japonês ocupam, respetivamente, o sexto e sétimo lugares e o urdu fica em último lugar. O teste de Kruskal Wallis revela que se mantêm diferenças significativas entre os valores.

Tabela 5.6: Análise da *informativeness* sem aplicação de *stemming*

		Mediana	IIQ
Inglês	en	12,38	19,47*
Chinês	zh	3,72	7,44
Árabe	ar	3,53	2,40
Francês	fr	3,01	3,20
Italiano	it	2,57	1,83
Hebreu	he	2,01	1,03
Japonês	ja	1,88	5,13
Catalão	ca	1,83	1,78
Russo	ru	1,68	1,34
Alemão	de	1,67	1,88
Português	pt	1,65	1,84
Bengali	bn	1,45	2,87
Coreano	ko	1,38	0,90
Grego	el	1,25	1,37
Indonésio	id	1,12	1,14
Persa	fa	1,10	1,33
Turco	tr	1,07	0,81
Hindi	hi	0,82	1,61
Urdu	ur	0,79	1,13
χ^2		4360,3	
Valor-p		<2,2e-16	

* valores diferentes daqueles verificados na Tabela 5.5

5.2.2.5 Consistency

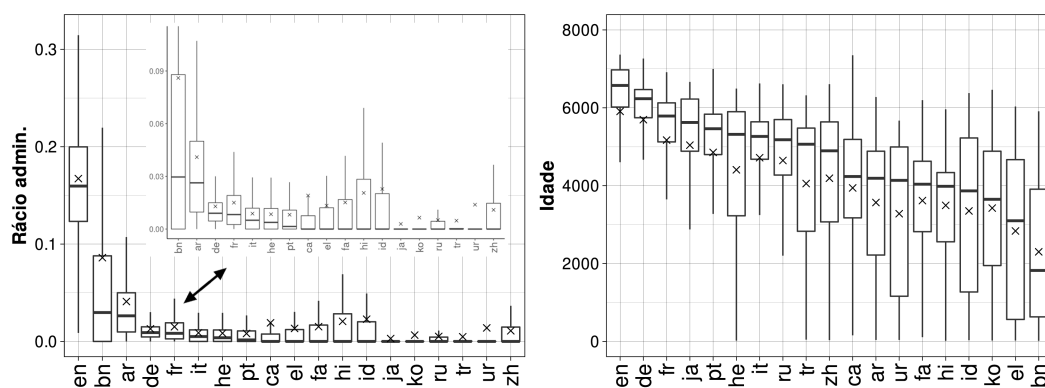
Na Tabela 5.7 podemos observar os resultados da métrica *consistency*. Da sua análise, salientam-se os idiomas com melhor no topo da classificação, por ordem decrescente: inglês, alemão e francês, novamente. O inglês volta, com a maior mediana volta a ser também o idioma com maior IIQ. O bengali ocupa outra vez a última posição. De salientar ainda, o facto da distribuição de valores ser mais homogénea nesta métrica do que nas anteriores, sendo que o idioma melhor classificado tem uma mediana menos de quatro vezes superior àquele que ocupa a última posição. Nesta métrica, os *outliers* correspondem, sobretudo, a artigos recentes e portanto sem nenhuma edição por administradores, resultando num rácio de edições por administradores nulo. O teste de Kruskal-Wallis revela, novamente, diferenças significativas entre os idiomas. Os testes de Dunn, revelam, no entanto, que não há diferenças significativas entre os dois primeiros classificados – inglês e alemão. também revelam que não existem diferenças significativas entre o último e o penúltimo classificados – bengali e grego, respetivamente.

Tabela 5.7: Avaliação da qualidade dos idiomas para a métrica *consistency*

		Mediana	IIQ	Idiomas significativamente inferiores																#	
Inglês	en	3286,4	476,6	fr	ja	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn	17
Alemão	de	3116,5	362,0	fr	ja	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn	17
Francês	fr	2893,5	501,2	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn			15
Japonês	ja	2811,5	669,5	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn			15
Português	pt	2731,0	518,5	he	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn					13
Hebreu	he	2659,0	1336,5	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn							11
Italiano	it	2632,0	480,3	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn							11
Russo	ru	2590,0	712,7	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn							11
Turco	tr	2532,5	1324,9	ar	ur	fa	hi	id	ko	el	bn										8
Chinês	zh	2447,0	1284,0	ca	ar	ur	fa	hi	id	ko	el	bn									9
Catalão	ca	2117,0	1007,0	ar	ur	fa	hi	ko	el	bn											7
Árabe	ar	2095,0	1333,5	el	bn																2
Urdu	ur	2067,8	1917,5	bn																	1
Persa	fa	2018,8	904,9	el	bn																2
Hindi	hi	1991,8	889,0	bn																	1
Indonésio	id	1933,0	1977,9	el	bn																2
Coreano	ko	1824,3	1467,6	el	bn																2
Grego	el	1549,0	2052,0																		0
Bengali	bn	911,3	1639,8																		0
χ^2		4534,3																			
Valor- <i>p</i>		<2,2e-16																			

* nível de significância $0,01 < p \leq 0,05$, nível de significância $p \leq 0,01$ para os restantes valores, IIQ: Intervalo interquartil, #: total de idiomas significativamente inferiores

Fazendo uma análise das medidas constituintes da *consistency*, de acordo com a Figura 5.5, verifica-se que o inglês está, como seria de esperar, em primeiro lugar na idade, sendo que corresponde à versão mais antiga da Wikipédia. O alemão, que tem a segunda versão mais antiga da Wikipédia, ocupa o segundo lugar. Apesar da terceira versão mais antiga ser o catalão, este ocupa a 11^a posição. A terceira posição é ocupada pelo francês e o último lugar cabe ao bengali. Quanto ao rácio de edições por administradores, o domínio do inglês é claro, aparecendo, em seguida, o bengali e o árabe. Observa-se que o rácio de edições feitas por administradores é geralmente baixo para todos os idiomas. Para 11 dos idiomas, o valor da mediana é 0, apesar da média ser superior. Verifica-se uma grande dispersão na medida da idade dos artigos para a generalidade dos idiomas.

Figura 5.5: Distribuição das medidas da métrica *consistency*

5.2.2.6 Volatility

Analisando a Tabela 5.8, observa-se que esta métrica não segue a distribuição habitual das anteriores. Tendo em conta que, um tempo de reversão médio menor equivalerá a uma recuperação de edições erradas, na *volatility*, uma mediana inferior corresponde a maior qualidade. Os primeiros lugares são, assim, ocupados pelo bengali, catalão, indonésio, coreano e urdu, todos eles com medianas de 0. O inglês surge apenas na quinta posição, e o último lugar pertence ao japonês. A análise de valores de mediana 0, implica um cuidado extra, como descrito na Secção 5.1. Cruzando estes dados com o número de reversões, observa-se que, no inglês, existem 0,3% de artigos sem qualquer reversão, enquanto que este valor sobe para os 36%, 23%, 36%, 28%, e 74%, no caso do bengali, catalão, indonésio, coreano e urdu, respetivamente. Verifica-se, ainda, que o urdu tem um IIQ de 0, uma vez que os poucos valores que tem são classificados como *outliers*. Tendo em conta estes aspectos, considera-se que estes cinco idiomas não alcançam, na verdade, valores de qualidade superiores nesta métrica. O primeiro lugar caberá, assim, ao inglês, seguido do francês. O teste de Kruskal-Wallis volta a revelar diferenças significativas entre os idiomas. Não se observam, contudo, diferenças significativas para os dois últimos idiomas da tabela – árabe e japonês. A única medida da *volatility* é o tempo médio de reversão das edições, e a sua dispersão está representada na Figura 5.6. Na observação da mesma confirmam-se os dados analisados na Tabela 5.8.

Tabela 5.8: Avaliação da qualidade dos idiomas para a métrica *volatility*

		Mediana	IIQ	Idiomas significativamente inferiores															#	
Bengali	bn	0	63,0	ur	en*	pt*	it	he	hi	ru	fa	tr	el	zh	ar	ja			13	
Catalão	ca	0	26,0	id*	ko	ur	en*	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja	16
Indonésio	id	0	50,3	ur	en	fr	it	he	hi	ru	fa	tr	el	zh	ar	ja				13
Coreano	ko	0	32,0	ur	en	fr	it*	he	hi	ru	fa	tr	el	zh	ar	ja				13
Urdu	ur	0	0,0	fr	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja				13
Inglês	en	2	3,0	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja					12
Francês	fr	4	6,0	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja					12
Português	pt	5	18,0	he	hi	ru	fa	tr	el	zh	ar	ja								9
Alemão	de	5,5	12,0	it	he	hi	ru	fa	tr	el	zh	ar	ja							10
Italiano	it	6,5	30,0	ru	fa	tr	el	zh	ar	ja										7
Hebreu	he	13	31,0	ru	fa	tr	el	zh	ar	ja										7
Hindi	hi	19	31,8	ru	fa	tr	el	zh	ar	ja										7
Russo	ru	26	77,5	ar	ja															2
Persa	fa	32,5	87,5	ar	ja															2
Turco	tr	33	99,0	ar*	ja															2
Grego	el	36	63,0	ar	ja															2
Chinês	zh	40	58,0	ja																1
Árabe	ar	56	83,0																	0
Japonês	ja	58	90,5																	0
χ^2		2775,9																		
Valor-p		<2,2e-16																		

* nível de significância $0,01 < p \leq 0,05$, nível de significância $p \leq 0,01$ para os restantes valores, IIQ: Intervalo interquartil, #: total de idiomas significativamente inferiores

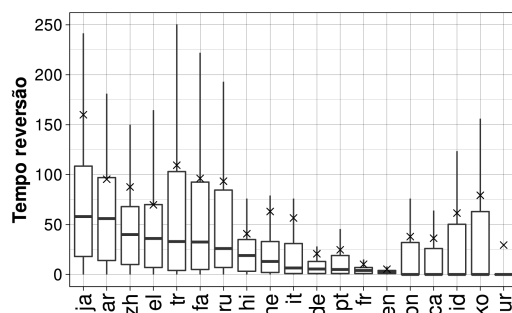


Figura 5.6: Distribuição da medida do tempo médio de reversão das edições

5.2.2.7 Currency

Na *currency*, uma mediana menor equivale a uma qualidade maior, dado que, um valor de atualidade pequeno corresponde a um artigo recentemente atualizado. Partindo deste princípio, constata-se a partir da Tabela 5.9 que o inglês ocupa novamente a primeira posição, seguido pelo alemão e pelo japonês. O inglês obtém novamente o maior IIQ entre os idiomas. O hindi e o urdu ocupam as duas últimas posições. O urdu, último classificado tem uma mediana 1.785% superior à da inglês. Os *outliers* correspondem, maioritariamente, a artigos onde já não é feita uma edição dos mesmos há muito tempo. O teste de Kruskal-Wallis volta a revelar diferenças significativas entre os idiomas. O alemão, não é, no entanto, significativamente diferente do japonês. O urdu, por outro lado, é significativamente diferente dos restantes idiomas. A única medida da *currency* é a atualidade e a dispersão dos seus valores pode ser observada na Figura 5.7. O diagrama de caixa ilustra os resultados analisados da Tabela 5.9.

Tabela 5.9: Avaliação da qualidade dos idiomas para a métrica *currency*

		Mediana	IIQ	Idiomas significativamente inferiores																#		
Inglês	en	20	17	de	ja	ca	ru	fr	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur	18
Alemão	de	42	72	ca	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur					14
Japonês	ja	42	39	ca	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur					14
Catalão	ca	50	99	ar	zh	bn	ko	el	id	pt	hi	ur										9
Russo	ru	50	90	fa	tr	it*	ar	zh	he	bn	ko	el	id	pt	hi	ur						13
Francês	fr	54	82	fa	tr	it*	ar	zh	he	bn	ko	el	id	pt	hi	ur						13
Persa	fa	57	114	bn	ko	el	id	pt	hi	ur												7
Turco	tr	63	66	bn	ko	el	id	pt	hi	ur												7
Italiano	it	65	101	bn	ko	el	id	pt	hi	ur												7
Árabe	ar	71	144	ko	el	id	pt	hi	ur													6
Chinês	zh	75	124	he	ko	el	id	pt	hi	ur												7
Hebreu	he	79	105	bn*	ko	el	id	pt	hi	ur												7
Bengali	bn	87	205	el	id	pt	hi	ur														5
Coreano	ko	126	169	el	id	pt	hi	ur														5
Grego	el	130	256	hi	ur																	2
Indonésio	id	140	349	hi	ur																	2
Português	pt	157	283	hi	ur																	2
Hindi	hi	210	218																			0
Urdu	ur	357	370																			0
χ^2		2802,9																				
Valor- <i>p</i>		<2,2e-16																				

* nível de significância $0,01 < p \leq 0,05$, nível de significância $p \leq 0,01$ para os restantes valores, IIQ: Intervalo interquartil, #: total de idiomas significativamente inferiores

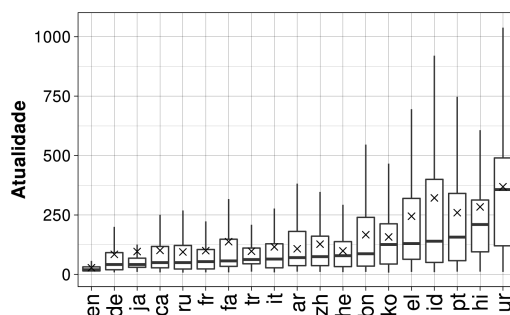


Figura 5.7: Distribuição da medida atualidade

5.2.3 Discussão de resultados

Quando analisamos as pontuações dos idiomas nas diferentes métricas e suas medidas, vemos que alguns idiomas geralmente ocupam os primeiros lugares e outros ocupam frequentemente os lugares inferiores. Para poder definir uma classificação de idiomas, é calculada a média do número de idiomas significativamente mais baixos em todas as métricas. Esses valores estão presentes na Tabela 5.10, onde os idiomas são classificados em ordem decrescente da média, uma vez que uma média mais alta significa que o idioma teve pontuação significativamente mais alta do que os restantes, logo, maior qualidade. Também estão representados os percentis de cada idioma. As melhores pontuações são destacadas em negrito. Na *currency*, é necessário ter em consideração as restrições associadas ao bengali, catalão, indonésio, coreano e urdu, descritas na Secção 5.2.2.6.

Como esperado, o inglês está no topo, com uma média de idiomas significativamente inferiores (ISI) de 14,0 e um percentil médio de 85%. O alemão e o francês obtiveram uma média de ISI de 12,0 e 11,7 e percentis médios de 80% e 77%, respetivamente. O último lugar pertence ao grego, com uma média de ISI significativamente mais baixa – 2,8 e um percentil médio de 25%. O urdu teve uma pontuação semelhante, com médias de 3,0 e 25%, para média de ISI e percentis. Dadas as particularidades da *volatility*, já discutidas, e considerando, assim, o inglês como o idioma melhor classificado, a única métrica em que este não se classifica em primeiro lugar é a *complexity*, mas essa métrica pode estar sujeita às restrições descritas anteriormente. Em relação aos idiomas selecionados pela sua tradição histórica, podemos observar que quatro deles – grego, persa, turco e coreano estão na metade inferior da tabela. Por outro lado, o italiano foi o idioma que obteve a melhor média dos idiomas significativamente inferiores – 9,0, com um percentil médio de 58%, seguido pelo catalão, com 8,5 e 64%.

Para entender se a qualidade está relacionada com a quantidade, é calculada a correlação entre a qualidade nas diferentes métricas e o número de falantes e o número total de artigos em cada versão da Wikipédia para os idiomas selecionados. Os resultados são mostrados na Tabela 5.11, que apresenta o valor de correlação de Spearman e os valores-*p* para o número de falantes e artigos da Wikipédia, para todos os idiomas, e os valores calculados da qualidade das métricas.

Analisando os resultados, verifica-se que existe uma correlação significativa entre a qualidade e o número de artigos totais em cada versão da Wikipédia, principalmente para *completeness*

Tabela 5.10: *Ranking* dos idiomas

	Authority		Completeness		Complexity		Informativeness		Consistency		Volatility		Currency		Mean	
	# ISI	%	# ISI	%	# ISI	%	# ISI	%	# ISI	%	# ISI	%	# ISI	%	ISI	%
Inglês	18	100%	18	100%	1	33%	14	100%	17	95%	12	68%	18	100%	14,0	85%
Alemão	17	95%	17	95%	3	67%	6	53%	17	95%	10	63%	14	89%	12,0	80%
Francês	13	79%	15	89%	2	50%	12	87%	15	84%	12	68%	13	79%	11,7	77%
Russo	14	84%	14	84%			6	53%	11	63%	2	21%	13	79%	10,0	64%
Italiano*	14	84%	13	74%	0	17%	11	80%	11	63%	7	42%	7	47%	9,0	58%
Catalão*	5	37%	6	53%			8	73%	7	47%	16	100%	9	74%	8,5	64%
Português	9	68%	11	68%	5	100%	6	53%	13	79%	9	58%	2	16%	7,9	63%
Hebreu*	6	53%	8	58%					11	63%	7	42%	7	47%	7,8	53%
Japonês	5	37%	5	37%					15	84%	0	5%	14	89%	7,8	51%
Árabe	5	37%	13	74%			13	93%	2	26%	0	5%	6	42%	6,5	46%
Chinês	12	74%	2	16%					9	58%	1	16%	7	47%	6,2	42%
Bengali	0	5%	2	16%			5	47%	0	5%	13	79%	5	32%	4,2	31%
Coreano*	1	21%	0	5%			3	33%	2	26%	13	79%	5	32%	4,0	33%
Turco*	6	53%	2	16%	3	67%	0	7%	8	53%	2	21%	7	47%	4,0	38%
Persa*	4	32%	5	37%			0	7%	2	26%	2	21%	7	47%	3,3	28%
Hindi	6	53%	5	37%			0	7%	1	16%	7	42%	0	5%	3,2	27%
Indonésio	0	5%	2	16%			0	7%	2	26%	13	79%	2	16%	3,2	25%
Urdu	1	21%	0	5%					1	16%	13	79%	0	5%	3,0	25%
Grego*	0	5%	9	63%			4	40%	0	5%	2	21%	2	16%	2,8	25%

ISI: Idiomas significativamente inferiores, * idiomas selecionados pela sua importância histórica

Tabela 5.11: Correlação entre as métricas e o número de falantes e de artigos

	Falantes		Wikipédia	
	correlação	valor- <i>p</i>	correlação	valor- <i>p</i>
Authority	0,67	0,0015*	0,9	0,0000**
Completeness	0,46	0,0459	0,94	0,0000**
Complexity	-0,11	0,8131	-0,13	0,7892
Informativeness	0,86	0,0000**	0,91	0,0000**
Consistency	0,29	0,2247	0,69	0,0011**
Volatility	0,01	0,9597	-0,15	0,5361
Currency	-0,08	0,7544	-0,46	0,0451

* nível de significância $p \leq 7,1e-3$, ** nível de significância $p \leq 1,4e-3$. (correção de Bonferroni para $p=0,05$ e $p=0,01$, 7 testes)

(0,94), *authority* (0,9) e *informativeness* (0,91), com valores-*p* significativos ($\leq 0,01$). *Informativeness* é a métrica com maior correlação (0,86) com o número de falantes, com um valor-*p* significativo ($\leq 0,01$), seguido de *authority*, com valor-*p* $\leq 0,05$. Há também uma forte correlação (0,63) para o número de falantes e o número de artigos em cada versão da Wikipédia, com um valor-*p* significativo ($\leq 0,05$).

5.3 Conclusão

Nesta primeira fase, foi comparada a qualidade de artigos de saúde da Wikipédia em 19 idiomas distintos, recorrendo a sete métricas propostas por Stvilia *et al.* [58]. Dos resultados, pode concluir-se que existem diferenças significativas na qualidade da informação em saúde entre os diferentes idiomas artigos disponíveis na Wikipédia. Dá-se, assim, resposta à primeira questão de investigação. O inglês foi o idioma que demonstrou maior diferença para os restantes idiomas, apresentando maior qualidade, seguido pelo alemão, francês e russo. O grego, o urdu, o indonésio e o hindi foram os idiomas com pior qualidade nas métricas avaliadas. Pode, ainda, concluir-se

que existe uma correlação da qualidade com a quantidade de falantes e o número de artigos disponíveis na Wikipédia. Esta correlação é mais significativa para o número de artigos da Wikipédia e, sobretudo, algumas métricas, como a *completeness*, *informativeness* ou *authority*.

O *dataset* resultante desta fase está disponível num repositório¹, em conjunto com o código desenvolvido. O *dataset* contém o resultado da avaliação de todos os artigos, em cada idioma, separados pelas diversas medidas e respetivas métricas e ordenados pela identificação única atribuída a cada artigo.

¹<https://doi.org/10.25747/ep0v-en19>

Capítulo 6

Proposta de medidas e métricas de avaliação específicas de saúde

Neste capítulo, será descrita a fase do trabalho relativa à proposta de medidas e consequentes métricas específicas de avaliação da qualidade da informação em artigos de saúde na Wikipédia. São propostas medidas de qualidade específicas, e analisadas, comparando com as medidas genéricas. Em seguida, aborda a proposta de métricas específicas de saúde, baseadas nas referidas medidas e sua avaliação.

6.1 Proposta de medidas específicas de saúde da Wikipédia

Para identificar medidas específicas, que possam ser utilizadas para avaliar a qualidade dos artigos da Wikipédia sobre saúde, foram analisados diversos artigos de diversas áreas da saúde e da medicina. Recolheram-se os elementos específicos comuns a esses artigos e procurou-se entender de que modo estes poderão contribuir para a avaliação da qualidade. Em seguida são descritas as medidas específicas propostas, resultantes dessa análise.

6.1.1 Número de *templates* de saúde

Os *templates* são elementos usados para estruturar informações na Wikipédia, permitindo que várias páginas reutilizem o mesmo elemento [85]. Simultaneamente, os *templates* permitem que os utilizadores tenham acesso rápido, fácil e organizado às informações. Os *templates* podem ser incluídos em qualquer área de um artigo da Wikipédia e são categorizados de acordo com seu assunto, em categorias de modelos e subcategorias dentro deles. Consideram-se apenas *templates* relacionados com saúde. A Figura 6.1 mostra um exemplo de um *template* relacionado com saúde – o *template* de classificações médicas – para a página Coronavírus 2019, com códigos médicos.

V · D · E	Pandemia de COVID-19	[Expandir]
V · D · E	Doenças causadas por vírus (A80-B34)	[Expandir]
Controle de autoridade	 Q84263196 · BNCf: 68698 · BNCHL: 000927635 · BNF: 17874453m · EBID: ID · GEC: 23470930 · GND: 1206347392 · LCCN: sh2020000570 · Treccani: COVID	
Identificadores	DiseasesDB: 60833 · DOID: D010080600 · eMedicine: 2500114 · ICD-10: U07.1 · KEGG: H02398 · MeSH: D000086382 · MeSH: C01.748.214 · NCIt: C171133	

Figura 6.1: *Template* do artigo sobre COVID-19 na Wikipédia lusófona. Fonte: https://pt.wikipedia.org/wiki/Pandemia_de_COVID-19

6.1.2 Número de valores nas *infoboxes* de saúde

As *infoboxes* são um tipo específico de *template* e um dos tipos mais recorrentes destes. São comumente utilizadas em artigos relacionados com a saúde. As *infoboxes* são uma tabela de formato fixo, geralmente disponível no canto superior direito das páginas. Estas contêm factos e estatísticas relevantes para os artigos relacionados com isto e melhoram a navegação entre eles. As *infoboxes* podem também incluir metadados. São uma forma de resumir aspectos importantes num formato fácil e rápido de ler para o utilizador. A Figura 6.2 representa um *infoboxes*, extraída da página sobre COVID-19 da Wikipédia portuguesa. Apenas foram recolhidas *infoboxes* relacionadas com a saúde. Estas contêm pares chave-valor, sendo as chaves previamente definidas para cada *infobox*, e os valores de cada uma delas podem variar. Existe também a possibilidade de incluir ou não cada um dos pares chave-valor, pelo que a contagem do número de pares pode indicar o grau de desenvolvimento e, portanto, a qualidade do artigo. Em artigos relacionados com a saúde, estes pares incluem frequentemente sintomas, complicações, tratamento e medicação, para doenças, ou dados farmacocinéticos, para medicação, como exemplos.

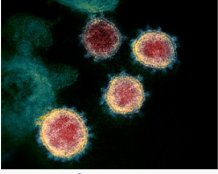
COVID-19	
	
Imagem de viriões de SARS-CoV-2 obtida por microscópio eletrónico de varrimento, em que se observam partículas virais a emergir de uma célula	
Sinónimos	Doença respiratória aguda por 2019-nCoV
Especialidade	Infectologia
Sintomas	Febre, tosse, falta de ar ^{[1][2]}
Complicações	Pneumonia, SDRA, sepsse, choque séptico, morte
Início habitual	2–14 dias após exposição ao vírus ^[1]
Causas	SARS-CoV-2
Fatores de risco	Idade avançada, doenças crónicas graves como doenças cardiovasculares, diabetes ou doenças pulmonares ^[1]
Método de diagnóstico	Exame PCR ^[4]
Prevenção	Lavar frequentemente as mãos com sabonete ou solução à base de álcool, tapar nariz e boca com o cotovelo ou lenço ao espirrar e tossir, evitar contacto próximo com pessoas infetadas, máscara cirúrgica, evitar sair de casa ^{[3][1]}
Tratamento	Sintomático e de apoio
Frequência	174 729 476 ^[1] casos confirmados desde dezembro de 2019
Mortes	3 768 560 ^[1] (2,21% dos casos confirmados) ^[3]
Classificação e recursos externos	
CID-10	U07.1 ^g
DiseasesDB	60833 ^g
eMedicine	2500114 ^g
MeSH	D000086382 ^g
 Leia o aviso médico	

Figura 6.2: *Infobox* do artigo sobre COVID-19 na Wikipédia lusófona. Fonte: https://pt.wikipedia.org/wiki/Pandemia_de_COVID-19

6.1.3 Número de imagens nas *infoboxes* de saúde

A Wikipédia permite a inclusão de elementos multimédia, como imagens. As imagens também podem ser incluídas em *infoboxes*, como mostra a Figura 6.2. Como elemento multimédia, as imagens enriquecem o conteúdo disponibilizado aos utilizadores, assumindo particular relevância em alguns temas como a saúde, onde procuram, por exemplo, os sinais de doenças, que são muitas vezes visuais. Neste caso, o seu número em *infoboxes*, categorizadas apenas no tópico relacionado com a saúde, é contabilizado.

6.1.4 Rácio de edições de administradores do WikiProject Medicine

No WikiProject Medicine, não existe qualquer requisito de formação para os seus membros. Ainda assim, a maioria deles são médicos, estudantes de medicina, enfermeiros, cientistas, e leigos com um interesse específico em certos tópicos médicos [90]. Assim, as edições dos administradores do WikiProject Medicine podem ser um indicador de qualidade em artigos relacionados com a saúde, potencialmente mais fiáveis do que o facto de serem apenas administradores gerais da Wikipédia.

6.1.5 Tradução pela Healthcare Translation Task Force

A Healthcare Translation Task Force foi criada como um projeto conjunto entre a WikiProject Medicine, a Wiki Project Med Foundation, e a Translators Without Borders. É um projeto baseado no voluntariado, contando desde 2019 com a ajuda de uma ferramenta de tradução. No início de 2021, já tinha mais de 1.900 artigos traduzidos em mais de 90 idiomas. A selecção dos artigos a traduzir pode ser um indicador de uma maior qualidade destes artigos.

6.1.6 Número de códigos médicos

Uma característica particular dos artigos relacionados com a saúde é a ligação com as classificações médicas. Os códigos médicos podem estar presentes nos diferentes *templates*, tal como o exemplo na Figura 6.1. As classificações médicas destinam-se a codificar diagnósticos ou procedimentos médicos. Um exemplo disso é a International Statistical Classification of Diseases and Related Health Problems – ICD [70]. Existem, no entanto, artigos de saúde onde os códigos médicos não são relevantes, como é o caso dos artigos sobre personalidades da área da saúde. Neste caso, apesar de não existirem códigos médicos, isso não se traduz, necessariamente, numa falta de qualidade do artigo.

6.1.7 Número de *links* reputados

Uma medida comumente utilizada para avaliar a qualidade da informação na Wikipédia são os *links* externos, garantindo a fiabilidade da informação. No entanto, o número destas ligações não é, por si só, uma garantia desta fiabilidade, uma vez que a autoridade não é garantida. Assim, propõe-se estimar a fiabilidade destes *links* utilizando a reputação das fontes de informação.

6.1.8 Número de secções recomendadas

O tamanho do artigo é uma medida usada frequentemente pelos autores para avaliar a qualidade de um artigo [6, 60, 58]. No entanto, quantidade não é sinónimo de qualidade. Para avaliar a qualidade semântica do texto, geralmente é feita uma validação manual, ou, por meio de ferramentas, que sempre implicam em uma avaliação manual, como é o caso da ferramenta DISCERN [44]. Como forma de avaliar automaticamente a semântica, ainda que de forma relativa, e ao mesmo tempo a correta estruturação do texto, propõe-se avaliar as diferentes secções dos artigos, verificando quais dessas estão na lista de secções recomendadas pelo WikiProject Medicine [82]. Esta medida pode assumir particular relevância na área da saúde, onde os temas, pela sua extensão, são propícios a esta estruturação. Este é o caso de um artigo sobre uma doença, onde deve haver secções para causas, sinais, sintomas, tratamento, epidemiologia, entre outros.

6.2 Análise de medidas de avaliação da qualidade

Nesta secção são analisadas as medidas de avaliação da qualidade em artigos de saúde da Wikipédia. Inicialmente é feita uma análise às medidas genéricas, já abordadas na fase I. Posteriormente, são analisadas as medidas específicas propostas na secção anterior, na avaliação da qualidade dos mesmos artigos.

6.2.1 Análise de medidas genéricas

Para analisar as medidas genéricas mais importantes na avaliação da qualidade, calcularam-se as medidas usadas por Stvilia *et al.* [58] no *dataset*. Como os dados não seguem uma distribuição normal, calculou-se a mediana de cada recurso no *dataset*. Em seguida, analisou-se a correlação de cada recurso com os níveis de qualidade. Para isso, utilizou-se o coeficiente de correlação de postos de Spearman, dado que os dados são ordinais. Converteu-se cada nível de qualidade num valor numérico, de 1 – Start a 5 – FA. A Tabela 6.1 mostra a mediana de cada medida, a sua correlação com a qualidade e o valor-*p* obtido num teste de hipóteses, com a hipótese nula de que a correlação é zero. Aplicou-se a correção de Bonferroni aos valores-*p* tendo em conta os vários testes de hipótese, indicando significância estatística. Apresentam-se as medidas em ordem decrescente de correlação com a qualidade.

A partir dos valores apresentados na Tabela 6.1, conclui-se que os valores de correlação variam desde uma correlação negligenciável de -0,04 para a “atualidade” até uma correlação moderada do “número de edições dos utilizadores registados” (0,53). Para a caracterização dos valores de correlação, adoptou-se a escala e terminologia utilizadas por Prion e Haerling [51]. Todas as medidas, excepto a “currency”, têm um valor de correlação significativamente diferente de 0. Entre estas, todas menos a “InfoNoise” são significativas para $p=0,01$.

Como esperado, o “Kincaid grade level” e a “diversidade” têm uma correlação negativa, uma vez que são pares subtrativos. A “atualidade” e o “tempo médio de reversão das edições” têm correlações negativas, uma vez que valores mais baixos correspondem a uma qualidade superior.

Tabela 6.1: Medianas de medidas genéricas e sua correlação com a qualidade

	Mediana	Correlação	valor-p
N.º de edições por utilizadores registados	1115,0	0,53	< 2,2e-16**
N.º de edições	1729,0	0,52	< 2,2e-16**
Conetividade	131,5	0,50	< 2,2e-16**
N.º de editores únicos	802,0	0,49	< 2,2e-16**
N.º de <i>links</i> externos	141,0	0,49	< 2,2e-16**
N.º de edições por utilizadores anónimos	550,5	0,47	< 2,2e-16**
N.º de reversões	148,5	0,47	< 2,2e-16**
Comprimento do artigo	24291,5	0,43	< 2,2e-16**
Idade do artigo	6726,5	0,38	< 2,2e-16**
N.º de imagens	13,0	0,37	< 2,2e-16**
Diversidade	0,5	-0,32	< 2,2e-16**
Rácio de edições por administradores	0,2	0,31	< 2,2e-16**
N.º de <i>links</i> internos	388,0	0,29	< 2,2e-16**
Tempo médio de reversão das edições	11,0	-0,28	< 2,2e-16**
Kincaid grade level	17,5	-0,25	8,473e-15**
Flesch reading ease	27,0	0,20	7,241e-10**
N.º de <i>links</i> internos quebrados	1,0	0,13	2,471e-5**
InfoNoise	0,88	0,13	5,989e-5*
Atualidade	6,0	-0,04	0,21

* significância $p < 2,6e-3$, ** significância $p < 5,3e-4$. (Correção de Bonferroni de $p=0,05$ e $p=0,01$, 19 testes)

Os resultados associados à “atualidade” podem estar relacionados com o facto de estarmos a trabalhar com os artigos mais vistos, com uma mediana de 6 dias. Note-se que no trabalho de Stvilia *et al.* [59] a mediana para artigos de nível FA era de três dias, enquanto que a mediana para um conjunto de artigos aleatórios era de 46 dias.

Analisando os valores das medianas, podem-se salientar algumas particularidades, tais como a elevada taxa de “número de edições por utilizadores registados” entre o total de edições. Outra particularidade que vale a pena mencionar é o “número de editores únicos”. Stvilia *et al.* [59] calcularam uma mediana de 108 editores únicos para o conjunto de níveis de FA e apenas cinco para o conjunto aleatório; neste *dataset*, calculou-se uma mediana de 802 editores únicos.

6.2.2 Análise de medidas específicas de saúde

Para investigar a importância das medidas específicas, propostas na Secção 6.1, na avaliação da qualidade, realizou-se uma análise semelhante à descrita na secção anterior, calculando a mediana de cada medida e a correlação de Spearman com a qualidade. Os resultados são mostrados na Tabela 6.2, ordenados de forma decrescente por valor de correlação. Além disso, realizou-se uma análise da distribuição de cada medida por nível de avaliação. Essas distribuições são mostradas na forma de diagramas de caixa, na Figura 6.3, onde um “X” representa a média.

A análise de correlação mostra que os valores são relativamente homogêneos, variando desde o mínimo de 0,19 para o “número de *imagens nas infoboxes*”, até ao máximo de 0,52 para o “número de *links* reputados”, uma correlação moderada. O “número de *links* reputados” tem um valor de correlação equivalente ao da segunda característica mais correlacionada (“número de edições”) das medidas genéricas, conforme descrito na Tabela 6.1. Note-se que os “códigos médicos” e a

Tabela 6.2: Medianas das medidas específicas e sua correlação com a qualidade

	Mediana	Correlação	valor- <i>p</i>
N.º de <i>links</i> reputados	46	0,52	< 2,2e-16**
N.º de secções recomendadas	5	0,40	< 2,2e-16**
Tradução pela Healthcare Translation Task Force	0	0,32	< 2,2e-16**
Rácio de edições de admin. do Wikiproject Medicine	0,05	0,25	1,9e-15**
N.º de <i>templates</i> de saúde	3	0,23	4,2e-13**
N.º de códigos médicos	0	0,23	1,1e-12**
N.º de valores nas <i>infoboxes</i> de saúde	8	0,21	8,2e-11**
N.º de imagens nas <i>infoboxes</i> de saúde	1	0,19	9,6e-10**

* significância $p < 6,3e-3$, ** significância $p < 1,3e-3$. (Correção de Bonferroni de $p=0,05$ e $p=0,01$, 8 testes)

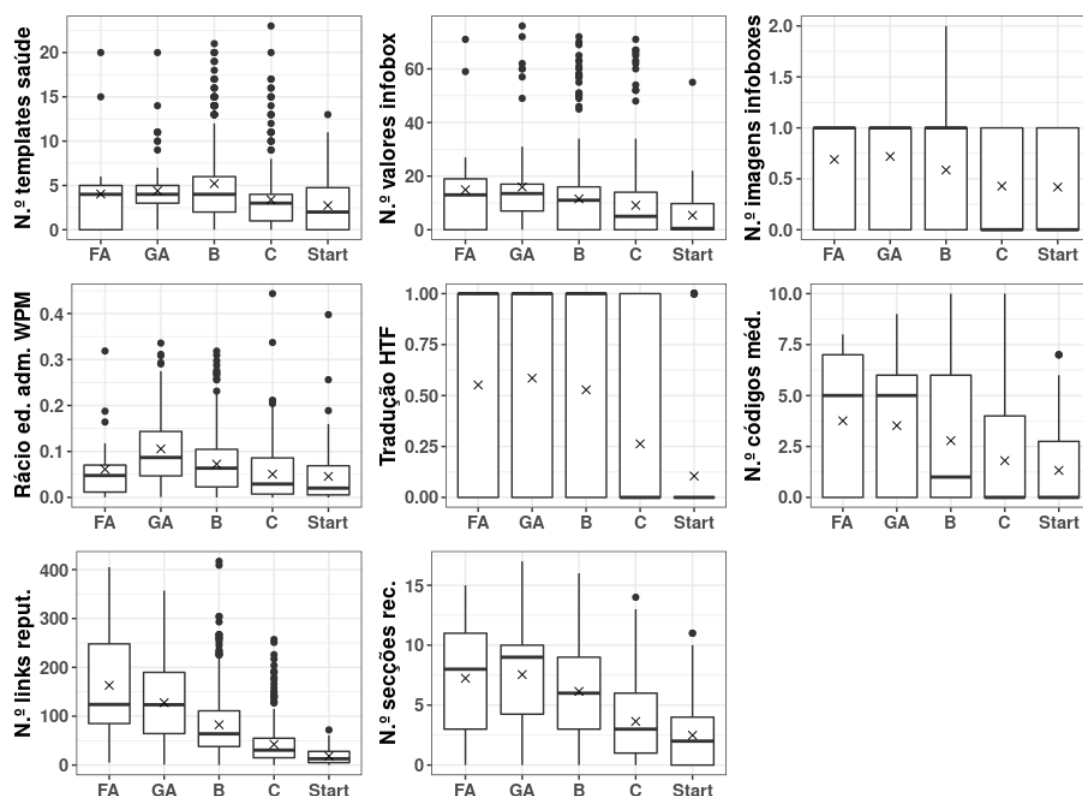


Figura 6.3: Distribuição das medidas específicas de saúde por nível de qualidade

“tradução pela Healthcare Translation Task Force” têm um valor mediano de 0, uma vez que a maioria dos artigos não estão na lista de artigos traduzidos e não têm códigos médicos. Todas as medidas têm um valor de correlação significativamente diferente de 0 com $p=0,01$.

Os diagramas de caixa permitem uma análise mais abrangente às diferenças por nível de qualidade. Em geral, os valores médios mostrados nos diagramas de caixa tendem a diminuir à medida que a qualidade aumenta. Existem algumas exceções, como é o caso do “Rácio de edições de administradores do Wikiproject Medicine”, entre os dois primeiros níveis – FA e GA. Isto pode indicar que estes dois tipos de artigos são muito semelhantes entre si, pelo que a distinção é difícil de fazer. Por outro lado, podemos notar o facto de que o último nível de qualidade – Start, é aquele

que geralmente se distingue mais dos restantes. Na análise dos diagramas de caixa, dois deles diferem dos restantes, relacionados com o “Num. imagens nas *infoboxes* de saúde” e o “Tradução”. O primeiro caso deve-se à pouca variação dos valores – de zero ao máximo de duas imagens e o segundo é causado pela natureza binária da variável – 1 se traduzida, 0 se não traduzida. Nestes casos, a média, representada no gráfico, fornece uma imagem mais clara da tendência. Podemos ver que existe uma grande diferença nos dois últimos níveis C, e Start – relativamente a estes aspectos, influenciando estes resultados.

6.3 Proposta de métricas específicas

Nesta secção são abordadas as métricas específicas que podem ser utilizadas para avaliar a qualidade dos artigos da Wikipédia relacionados com saúde. Inicialmente é analisado a importância das medidas genéricas, de modo a orientar a construção das métricas específicas. Posteriormente, são propostas as métricas específicas, recorrendo a medidas genéricas e medidas específicas.

6.3.1 Importância das medidas nas métricas genéricas

Calculou-se a correlação de Pearson entre cada métrica e as respetivas medidas para determinar a contribuição geral de cada medida para o valor final de cada métrica. A Tabela 6.3 mostra as medidas organizados por métrica, com os valores de correlação e o valor-*p* obtido num teste de hipóteses, sendo a hipótese nula de que a correlação é zero.

Tabela 6.3: Correlação das métricas com as suas medidas

	Correlação	valor- <i>p</i>	Métrica
N.º de edições	0,99	< 2,2e-16**	Authority
N.º de editores únicos	0,95	< 2,2e-16**	
N.º de edições por ut. registados	0,93	< 2,2e-16**	
N.º de edições por ut. anónimos	0,92	< 2,2e-16**	
N.º de reversões	0,89	< 2,2e-16**	
N.º de <i>links</i> externos	0,60	< 2,2e-16**	
Conetividade	0,28	< 2,2e-16**	Completeness
Comprimento do artigo	1,00	< 2,2e-16**	
N.º de <i>links</i> internos	0,31	< 2,2e-16**	
Nº de <i>links</i> int. quebrados	0,02	0,49	Complexity
Flesch	1,00	< 2,2e-16**	
Kincaid	-0,95	< 2,2e-16**	Informativeness
N.º de imagens	1,00	< 2,2e-16**	
Diversidade	-0,27	< 2,2e-16**	
InfoNoise	0,08	1,1e-2*	Consistency
Idade	1,00	< 2,2e-16**	
Rácio de edições por administradores	0,17	7,31e-8**	Currency
Atualidade	1,00	< 2,2e-16**	Volatility
Tempo médio de reversão	1,00	5,3e-2	

* respetivos níveis de significância: $p < [7,1e-3, 1,6e-2, 2,5e-2, 1,6e-2, 2,5e-2, 5e-2, 5e-2]$, ** respetivos níveis de significância: $p < [1,4e-3, 3,3e-3, 5e-3, 3,3e-3, 5e-3, 1e-2, 1e-2]$. (Correção de Bonferroni para $p=0,05$ e $p=0,01$, [7, 3, 2, 3, 2, 1, 1] testes respetivos)

Observa-se uma heterogeneidade significativa nos valores de correlação. Existem medidas com uma correlação muito forte, incluindo valores de 1, mostrando uma alta contribuição para o valor final da métrica. Destas correlações fortes, apenas a correlação do “tempo de reversão médio” com a “volatilidade” não é significativamente diferente de 0.

Por outro lado, existem correlações desprezíveis, como o “número de *links* internos quebrados” (0,02) e “InfoNoise” (0,08). O primeiro não é significativamente diferente de 0, mas o “InfoNoise” é significativamente diferente de 0 com um valor-*p* de 0,05.

A consistência representa um caso notável, pois as duas medidas constituintes alcançaram resultados diferentes; a medida de “idade” teve uma correlação muito forte (1), enquanto o “rácio das edições por administradores” teve um valor insignificante de 0,17. Para a “atualidade” e “volatilidade”, as medidas tiveram o valor 1, provavelmente influenciado por ter apenas uma medida. No entanto, a correlação do “tempo médio de reversão” com a “volatilidade” não é significativamente diferente de 0.

6.3.2 Proposta de métricas específicas de saúde

Para propor métricas específicas que possam ser usadas para avaliar a qualidade dos artigos da Wikipédia relacionados com saúde, adaptaram-se as métricas genéricas de Stvilia *et al.* [58], adicionando ou substituindo medidas.

Para obter o peso de cada medida, considerou-se a sua importância e especificidade para avaliar a qualidade da informação em saúde. Também se considerou o peso das medidas respetivas de cada métrica genérica, atribuindo o peso de cada medida proposta, de acordo com o valor mediano calculado, de modo a que o resultado final para aquela métrica corresponda ao mesmo intervalo de valores dos existentes. Por exemplo, o “rácio de edições por administradores” teve uma mediana de 0,2 e um peso de 0,6; o seu substituto, o “rácio de edições por administradores do Wikiproject Medicine” teve uma mediana inferior – 0,05, pelo que o peso aumenta para 1,9.

Complexity, *currency* e *volatility* permaneceram inalteradas, uma vez que nenhuma das medidas propostas se enquadra nessas métricas. As restantes são propostas da seguinte forma:

$$\textbf{HealthAuthority} = 0,2 * N.^{\circ} \text{ de editores \u00fanicos} + 0,2 * N.^{\circ} \text{ de edi\u00e7\u00f5es} + 0,1 * \text{Conetividade} + 0,3 * N.^{\circ} \text{ de revers\u00f5es} + 0,6 * N.^{\circ} \text{ de Links Reputados} + 0,1 * N.^{\circ} \text{ de edi\u00e7\u00f5es por utilizadores registados} + 0,2 * N.^{\circ} \text{ de edi\u00e7\u00f5es por utilizadores an\u00f3nimos} + 200 * \text{Tradu\u00e7\u00e3o pela HTF}$$

$$\textbf{HealthCompleteness} = 0,4 * N.^{\circ} \text{ de links internos quebrados} + 0,4 * N.^{\circ} \text{ de links internos} + 0,2 * \text{Comprimento do artigo} + 970 * N.^{\circ} \text{ de sec\u00e7\u00f5es recomendadas} + 4850 * N.^{\circ} \text{ de c\u00f3digos m\u00e9dicos}$$

$$\textbf{HealthInformativeness} = 0,6 * \text{InfoNoise} - 0,6 * \text{Diversidade} + 0,3 * N.^{\circ} \text{ de Imagens} + N.^{\circ} \text{ de valores nas Infoboxes de sa\u00fade} + N.^{\circ} \text{ de imagens nas Infoboxes de sa\u00fade} + 0,08 * N.^{\circ} \text{ Templates de sa\u00fade}$$

$$\text{HealthConsistency} = 1,9 * \text{Rácio de edições por administradores do WikiProject Medicine} + 0,5 * \text{Idade}$$

A “tradução pela HTF” foi adicionada à HealthAuthority, pois a seleção para tradução pela Task Force pode reforçar a autoridade desse artigo. Os “*links* reputados” substituíram os “*links* externos” para filtrar os *links* externos pela sua reputação na área da saúde. Em HealthCompleteness, foi adicionado o “número de secções recomendadas” e o “número de códigos médicos”, pois ambas as medidas podem indicar um tamanho adequado da informação. Em HealthInformativeness, foi adicionado o “número de imagens”, os “valores em *infoboxes* de saúde” e o “número de *templates* médicos”, pois abordam a quantidade de informação num artigo. Para HealthConsistency, o “rácio de edições por administradores” foi substituído pelo “rácio de edições por administradores do WPM”, representando administradores relacionados com a saúde.

6.4 Avaliação das métricas de avaliação

Para avaliar métricas específicas e compará-las a métricas genéricas no domínio da saúde, calculou-se a correlação de ambos os tipos de métricas com os níveis de qualidade da Wikipédia em cada dimensão de análise: autoridade, integridade, informatividade e consistência. Também se conduziram testes estatísticos para comparar as duas correlações para cada dimensão. Os valores de correlação e os valores-*p* associados às comparações são apresentados na Tabela 6.4.

Tabela 6.4: Correlação entre métricas genéricas e específicas

	Genérica	Específica	valor- <i>p</i>
(Health)Authority	0,43	0,46	0**
(Health)Completeness	0,34	0,36	0,58
(Health)Informativeness	0,13	0,23	9,0e-4**
(Health)Consistency	0,30	0,30	1,0

* nível de significância $p < 0,05$, ** nível de significância $p < 0,01$.

A partir da análise da tabela, podemos concluir que se melhoraram todas as métricas, embora com diferenças heterogéneas. A métrica que mostrou as melhorias mais marginais (arredondada para zero) foi HealthConsistency. Este resultado está provavelmente associado ao fato de que o peso da medida – Rácio de edições por administradores do WikiProject Medicine – é muito insignificante no total da fórmula. No outro extremo, HealthInformativeness, representa uma melhoria muito significativa, revelando a provável importância dos *templates* e suas características na avaliação da qualidade dos artigos da Wikipédia.

6.5 Conclusão

Nesta fase, foram propostas medidas específicas para avaliação da qualidade dos artigos em saúde na Wikipédia. Depois de uma análise às mesmas, concluiu-se que as mais eficazes são o número de *links* reputados, o número de secções recomendadas e a tradução pela Healthcare Translation Task Force. Foram, posteriormente, propostas métricas de avaliação específicas para artigos de saúde, baseadas nas métricas genéricas em conjugação com as medidas específicas. Essas métricas foram a HealthAuthority, HealthCompleteness, HealthInformativeness e HealthConsistency. A HealthInformativeness e HealthAuthority apresentaram os melhores resultados. Conclui-se, portanto, que é possível melhorar a avaliação da qualidade dos artigos de saúde da Wikipédia utilizando métricas específicas, respondendo, assim, à segunda questão de investigação.

O *dataset* resultante desta fase está disponível num repositório institucional¹. O *dataset* contém o resultado da avaliação dos artigos, separados pelas diversas medidas e respetivas métricas, genéricas e específicas, e ordenados pela identificação única atribuída a cada artigo.

¹<https://doi.org/10.25747/wfzk-h937>

Capítulo 7

Conclusões e trabalho futuro

Neste capítulo final, conclui-se sobre o cumprimento dos objetivos propostos, comentando os resultados e os desafios encontrados. São ainda tecidas sugestões para trabalho futuro.

7.1 Conclusões

Pela importância da Wikipédia enquanto fonte de informação, em particular na saúde, surge a necessidade de avaliar a qualidade da informação. Sendo uma plataforma multi-idioma, um dos objetivos desta dissertação é avaliar a equidade no acesso à informação garantida pela Wikipédia em diversos idiomas. O segundo objetivo, parte da identificação de especificidades nos artigos de saúde da Wikipédia, pretendendo-se avaliar se é possível, a partir destas, melhorar as métricas já existentes de avaliação automática da qualidade da informação em saúde, na Wikipédia.

Numa primeira fase, foi feita uma comparação entre artigos sobre saúde e medicina da Wikipédia em 19 idiomas diferentes: inglês, árabe, francês, português, alemão, persa, italiano, chinês, russo, japonês, hebreu, coreano, catalão, indonésio, turco, grego, hindi, bengali e urdu. Para avaliar a qualidade dos artigos, foram utilizadas sete métricas pré-definidas: *authority*, *completeness*, *complexity*, *informativeness*, *consistency*, *currency*, e *volatility*. Depois da análise dos resultados, concluiu-se que existem diferenças significativas na qualidade da informação em saúde nos artigos disponíveis na Wikipédia em diferentes idiomas, respondendo assim à primeira questão de investigação. O inglês foi o idioma que demonstrou maior qualidade, e o grego foi o idioma com pior qualidade na generalidade das métricas avaliadas. Concluiu-se, ainda, que existe uma correlação da qualidade com a quantidade de falantes e o número de artigos disponíveis na Wikipédia. Para combater estas assimetrias, a Wikimedia Foundation tem um projeto em curso – Abstract Wikipedia [71]. Este projeto visa criar uma versão da Wikipédia independente do idioma, modelando dados do Wikidata. Isso permitirá que as pessoas criem conteúdo independente do idioma que será posteriormente traduzido por meio de *software*. Este projeto também contém o Wikifunctions, que inclui o código que converte o conteúdo da Abstract Wikipedia para a Wikipédia no idioma pretendido.

Numa segunda fase, foram exploradas medidas específicas para avaliar artigos da Wikipédia sobre saúde. Assim, foram propostas oito medidas: número de *templates* de medicina, número de valores em *infoboxes* de saúde, número de imagens em *infoboxes* de saúde, rácio de edições por administradores do WikiProject Medicine, tradução pela Healthcare Translation Task Force, número de códigos médicos, número de *links* reputados e número de secções recomendadas. Destas, concluiu-se que as mais eficazes na avaliação da qualidade são o número de *links* reputados, o número de secções recomendadas e a tradução pela Healthcare Translation Task Force. Nas medidas genéricas, as mais eficazes são as edições feitas por utilizadores registados, conectividade e o número de edições. Tendo isto em conta, foram propostas métricas de avaliação específicas para artigos de saúde, baseadas nas métricas genéricas. Essas métricas foram a HealthAuthority, HealthCompleteness, HealthInformativeness e HealthConsistency. A HealthInformativeness e HealthAuthority mostraram melhorias estatisticamente significativas. Responde-se, assim, à segunda questão de investigação, concluindo que é possível melhorar a avaliação da qualidade dos artigos de saúde da Wikipédia utilizando métricas específicas.

Além dos desafios inerentes à grande quantidade de dados, surgiram desafios ligados à heterogeneidade dos idiomas analisados. Esta heterogeneidade foi particularmente relevante na aplicação de medidas e métricas que não foram desenvolvidas para lidar com especificidades dos vários idiomas, ou foram mesmo desenvolvidas para um idioma em particular, como é o caso dos índices de inteligibilidade. A par da heterogeneidade dos idiomas, surge a heterogeneidade das versões da Wikipédia associadas a cada idioma. Além das questões inerentes ao idioma, somam-se diferenças na estrutura de cada versão da Wikipédia, aliadas a regras e linhas orientadoras diferentes entre as mesmas. Tudo isto contribuiu para os grandes desafios levantados na recolha e posterior análise dos dados.

7.2 Trabalho futuro

Como trabalho futuro, sugere-se uma extensão do número de idiomas para comparação da qualidade. Sugere-se ainda uma análise mais profunda das medidas e métricas específicas para a saúde propostas. Isto deverá levar a criar novas medidas ou à alteração do peso atribuído a cada medida. Um passo importante será ainda a definição de métricas e medidas que possam substituir aquelas aplicáveis a apenas alguns idiomas, como é o caso dos índices de ilegibilidade ou do infoNoise. Outra sugestão é a aplicação das métricas específicas de saúde propostas para avaliação da qualidade de outros idiomas além do inglês. Nesse sentido, será necessário ter em conta que nem todas as versões da Wikipédia contam com um equivalente do WikiProject Medicine. Por outro lado, os projetos existentes têm graus de desenvolvimento muito diferentes, muitos deles muito incompletos, onde será impossível utilizar, por exemplo, as secções recomendadas. Será necessário ainda ter em conta que a estrutura dos mesmos é muito heterogênea, havendo grandes diferenças, por exemplo, ao nível da avaliação da qualidade dos artigos, quando disponível. Finalmente, estas métricas poderão ser utilizadas para uma avaliação automática da qualidade da informação, especificamente em artigos de saúde da Wikipédia.

Referências

- [1] Amazon. The top 500 sites on the web. Disponível em <https://www.alexa.com/topsites>, junho 2021.
- [2] Amin Azzam, David Bresler, Armando Leon, Lauren Maggio, Evans Whitaker, James Heilman, Jake Orlowitz, Valerie Swisher, Lane Rasberry, Kingsley Otoide, Fred Trotter, Will Ross, e Jack D. Mccue. Why medical schools should embrace wikipedia: Final-year medical student contributions to Wikipedia Articles for Academic Credit at One School. *Academic Medicine*, 92(2):194–200, 2017.
- [3] Rita Baeten, Slavina Spasova, Bart Vanhercke, e Stéphanie Coster. *Inequalities in access to healthcare*. November 2018.
- [4] Joshua E. Blumenstock. Size matters: Word count as a measure of quality on Wikipedia. *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, páginas 1095–1096, 2008.
- [5] Adam R. Brown. Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. *PS - Political Science and Politics*, 2(April):339–343, 2011.
- [6] Riccardo Conti, Emanuel Marzini, Angelo Spognardi, Ilaria Matteucci, Paolo Mori, e Marinella Petrocchi. Maturity assessment of Wikipedia medical articles. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, páginas 281–286, 2014.
- [7] Baptiste De La Robertie, Yoann Pitarch, e Olivier Teste. Measuring article quality in Wikipedia using the collaboration network. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, páginas 464–471, 2015.
- [8] Division for Sustainable Development Goals - United Nations Department of Economic and Social Affairs. Transforming our world: the 2030 Agenda for Sustainable Development. Disponível em <https://sustainabledevelopment.un.org/post2015/transformingourworld>, junho 2021.
- [9] Gil Domingues e Carla Teixeira Lopes. Characterizing and comparing Portuguese and English Wikipedia medicine-related articles. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, páginas 1203–1207, 2019.
- [10] Ateşman E. Measuring readability in turkish. *Tömer Language Journal*, 58:171–174, 1997.
- [11] Encyclopædia Britannica, inc. Jimmy Wales. Disponível em <https://global.britannica.com/biography/Jimmy-Wales>, junho 2021.

- [12] Ethnologue. How many languages are there in the world? Disponível em <https://www.ethnologue.com/guides/how-many-languages>, junho 2021.
- [13] Explosion. spaCy. Disponível em <https://spacy.io/>, 2021.
- [14] Viégas Fernanda, Wattenberg Martin, e Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. *Conference on Human Factors in Computing Systems - Proceedings*, 6(1):575–582, 2004.
- [15] Elena Filatova. Directions for exploiting asymmetries in multilingual wikipedia. Em *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, CLIAWS3 '09, página 30–37, USA, 2009. Association for Computational Linguistics.
- [16] R FLESCHE. A new readability yardstick. *The Journal of applied psychology*, 32(3):221–233, June 1948.
- [17] Python Software Foundation. bangla-stemmer 1.0. Disponível em <https://pypi.org/project/bangla-stemmer/>, 2021.
- [18] Python Software Foundation. Persianstemmer 1.0.0. Disponível em <https://pypi.org/project/PersianStemmer/>, 2021.
- [19] Python Software Foundation. snowballstemmer 2.1.0. Disponível em <https://pypi.org/project/snowballstemmer/>, 2021.
- [20] Python Software Foundation. textstat 0.7.0. Disponível em <https://pypi.org/project/textstat/>, 2021.
- [21] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–1, 01 2006.
- [22] Google. O ano em Pesquisa de 2020. Disponível em <https://trends.google.com/trends/yis/2020/GLOBAL/>, junho 2021.
- [23] Carol A. Haigh. Wikipedia as an evidence source for nursing and healthcare students. *Nurse Education Today*, 31(2):135–139, 2011.
- [24] Scott Hale. Multilinguals and wikipedia editing. *WebSci 2014 - Proceedings of the 2014 ACM Web Science Conference*, 12 2013.
- [25] James M. Heilman, Eckhard Kemmann, Michael Bonert, Anwesh Chatterjee, Brent Ragar, Graham M. Beards, David J. Iberri, Matthew Harvey, Brendan Thomas, Wouter Stomp, Michael F. Martone, Daniel J. Lodge, Andrea Vondracek, Jacob F. De Wolff, Casimir Liber, Samir C. Grover, Tim J. Vickers, Bertalan Meskó, e Michaël R. Laurent. Wikipedia: A key tool for global public health promotion. *Journal of Medical Internet Research*, 13(1):1–16, 2011.
- [26] James M Heilman e Andrew G West. Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language. *J Med Internet Res*, 17(3):e62, Mar 2015.
- [27] Imran Khan, Shahid Hussain, Hina Gul, Muhammad Shahid, e Muhammad Jamal. *An empirical study to predict the quality of wikipedia articles*, volume 932. Springer International Publishing, 2019.

- [28] J. Kincaid, R. Fishburn, e B. Chissom. Derivation of new readability formulas for navy enlisted personnel. 01 1975.
- [29] Marlene Kritz, Manfred Gschwandtner, Veronika Stefanov, Allan Hanbury, e Matthias Samwald. Utilization and perceived problems of online medical resources and search tools among different groups of european physicians. *Journal of medical Internet research*, 15:e122, 06 2013.
- [30] Larry Sanger. Larry Sanger - Education, Internet, and more. Disponível em <https://larrysanger.org/>, junho 2021.
- [31] Michaël R. Laurent e Tim J. Vickers. Seeking health information online: Does wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, 2009.
- [32] Yang W. Lee, Diane M. Strong, Beverly K. Kahn, e Richard Y. Wang. Aimq: a methodology for information quality assessment. *Information & Management*, 40(2):133–146, 2002.
- [33] Dirk Lewandowski e Ulrike Spree. Ranking of wikipedia articles in search engines revisited: Fair ranking for reasonable quality? *Journal of the American Society for Information Science and Technology*, 62(1):117–132, 2011.
- [34] Xinyi Li, Jintao Tang, Ting Wang, Zhunchen Luo, e Maarten de Rijke. Automatically assessing wikipedia article quality by exploiting article–editor networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9022:574–580, 2015.
- [35] Daniel A. London, Steven M. Andelman, Anthony V. Christiano, Joung Heon Kim, Michael R. Hausman, e Jaehon M. Kim. Is Wikipedia a complete and accurate source for musculoskeletal anatomy? *Surgical and Radiologic Anatomy*, 41(10):1187–1192, 2019.
- [36] Andrew Madden. A definition of information. *Aslib Proceedings*, 52:343–349, 11 2000.
- [37] Edison Marrese-Taylor, Pablo Loyola, e Yutaka Matsuo. An Edit-centric Approach for Wikipedia Article Quality Assessment. páginas 381–386, 2019.
- [38] T B F Martins, C M Ghiraldelo, Maria das Graças Volpe Nunes, e Osvaldo Novais de Oliveira Junior. Readability formulas applied to textbooks in brazilian portuguese, 1996.
- [39] Sorin Matei e Brian Britt. *Structural Differentiation in Social Media: Adhocracy, Entropy, and the “1% Effect”*. 01 2017.
- [40] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Nielsen, e Arto Lanamäki. "the sum of all human knowledge": A systematic review of scholarly research on the content of wikipedia. *Journal of the Association for Information Science and Technology*, 66, 02 2015.
- [41] Holmes Miller. The multiple dimensions of information quality. *Information Systems Management*, 13(2):79–82, 1996.
- [42] Omeed Modiri, Daipayan Guha, Naif M. Alotaibi, George M. Ibrahim, Nir Lipsman, e Aria Fallah. Readability and quality of wikipedia pages on neurosurgical topics. *Clinical Neurology and Neurosurgery*, 166(January):66–70, 2018.

- [43] V. Nanda. *Quality Management System Handbook for Product Development Companies*. Taylor & Francis, 2005.
- [44] NHS Executive Research and Development Programme. DISCERN online. Disponível em <http://www.discern.org.uk/index.php>, junho 2021.
- [45] Aoife O’Carroll, Erin Westby, Joseph Dooley, e Kevin Gordon. Information-seeking behaviors of medical students: A cross-sectional web-based survey. *JMIR Medical Education*, 1, 06 2015.
- [46] Elizabeth Park, Tiny Masupe, Joseph Joseph, Ari Ho-Foster, Afton Chavez, Swetha Jammalamadugu, Andrew Marek, Ruth Arumala, Dineo Ketshogileng, Ryan Littman-Quinn, e Carrie Kovarik. Information needs of botswana health care workers and perceptions of wikipedia. *International Journal of Medical Informatics*, 95:8–16, 2016.
- [47] Eunjeong L. Park e Sungzoon Cho. Konlpy: Korean natural language processing in python. Em *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Chuncheon, Korea, October 2014.
- [48] Daniel Pimienta, D. Prado, e Á Blanco. Twelve years of measuring linguistic diversity in the internet: balance and perspectives. *Paris: UNESCO. Retrieved March*, 7(September):2010, 2009.
- [49] Natalia Pletneva, Sarah Cruchet, Maria Ana Simonet, Maki Kajiwarra, e Célia Boyer. Results of the 10th hon survey on health and medical internet use. *Studies in Health Technology and Informatics*, 169(2008):73–77, 2011.
- [50] Lakha Prasannan, Nagaraj Gabbur, e Michele Haughton. Use of web resources among medical students at a large urban medical center. *Obstetrics and gynecology*, 123 Suppl 1:118S, 05 2014.
- [51] Susan Prion e Katie Haerling. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10:535–536, 10 2014.
- [52] Malolan S. Rajagopalan, Vineet K. Khanna, Yaacov Leiter, Meghan Stott, Timothy N. Showalter, Adam P. Dicker, e Yaacov R. Lawrence. Patient-Oriented Cancer Information on the Internet: A Comparison of Wikipedia and a Professionally Maintained Database. *Journal of Oncology Practice*, 7(5):319–323, 2011.
- [53] Azucena Santillan Garcia, Miguel Angel Mañez, e Xose Meijome. Uso de wikipedia por los profesionales de la salud. *Gaceta Sanitaria*, 28, 11 2014.
- [54] Pnina Shachaf e Noriko Hara. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3):357–370, 2010.
- [55] Thomas Shafee, Gwinyai Masukume, Lisa Kipersztok, Diptanshu Das, Mikael Häggström, e James Heilman. Evolution of wikipedia’s medical content: Past, present and future. *Journal of Epidemiology and Community Health*, 71(11):1122–1129, 2017.
- [56] Thomas Shafee, Gwinyai Masukume, Lisa Kipersztok, Diptanshu Das, Mikael Häggström, e James Heilman. Evolution of wikipedia’s medical content: Past, present and future. *Journal of Epidemiology and Community Health*, 71(11):1122–1129, 2017.

- [57] Andrea Simpson, Michelle Le, e Alicja N. Malicka. The Accuracy and Readability of Wikipedia Articles on Hearing Loss. *Journal of Consumer Health on the Internet*, 22(4):323–336, 2018.
- [58] B. Stvilia, M. B. Twidale, L. Gasser, e L. C. Smith. Information quality in a community-based encyclopedia. *Knowledge Management: Nurturing Culture, Innovation, and Technology-Proceedings of the 2005 International Conference on Knowledge Management*, páginas 101–113, 2005.
- [59] Besiki Stvilia, Michael Twidale, Linda Smith, e Les Gasser. Assessing information quality of a community-based encyclopedia. *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005*, 01 2005.
- [60] Athikhun Suwannakhan, Daniel Casanova-Martínez, Laphatrada Yurasakpong, Punchalee Montriwat, Krai Meemon, e Taweetham Limpanuparb. The Quality and Readability of English Wikipedia Anatomy Articles. *Anatomical Sciences Education*, 13:1–13, 2019.
- [61] Yu Suzuki e Masatoshi Yoshikawa. Assessing quality score of wikipedia articles using mutual evaluation of editors and texts. *International Conference on Information and Knowledge Management, Proceedings*, páginas 1727–1732, 2013.
- [62] Janice A. Taylor, Christiana M. Shaw, Sanda A. Tan, e John L. Falcone. Are the kids alright? Review books and the internet as the most common study resources for the general surgery clerkship. *American Journal of Surgery*, 215(1):191–195, jan 2018.
- [63] The Editors of Encyclopaedia Britannica. Cebuano language. Disponível em <https://www.britannica.com/topic/Cebuano-language>, junho 2021.
- [64] Garry R. Thomas, Lawson Eng, Jacob F. de Wolff, e Samir C. Grover. An Evaluation of Wikipedia as a Resource for Patient Education in Nephrology. *Seminars in Dialysis*, 26(2):159–163, 2013.
- [65] Neil Thompson e Douglas Hanley. Science is shaped by wikipedia: Evidence from a randomized control trial. *SSRN Electronic Journal*, 2017.
- [66] John Torous, Jamie Franzan, Ryan O'Connor, Ian Mathew, Matcheri Keshavan, Robert Kitts, e Robert Boland. Psychiatry residents' use of educational websites: A pilot survey study. *Academic Psychiatry*, 39, 06 2015.
- [67] UNESCO. A Decade of promoting multilingualism in cyberspace. Disponível em <http://en.unesco.kz/a-decade-of-promoting-multilingualism-in-cyberspace>, 2015.
- [68] United Nations University. Wikipedia Survey – Overview of Results. Disponível em https://www.ris.org/uploadi/editor/1305050082Wikipedia_Overview_15March2010-FINAL.pdf, junho 2021.
- [69] Ziko VanDijk. Wikipedia and lesser-resourced languages. *Language Problems and Language Planning*, 33(3):234–250, 2009.
- [70] WHO. International Statistical Classification of Diseases and Related Health Problems (ICD). Disponível em <https://www.who.int/standards/classifications/classification-of-diseases>, junho 2021.

- [71] Wikimedia. Abstract wikipedia. Disponível em https://meta.wikimedia.org/wiki/Abstract_Wikipedia, 2021.
- [72] Wikimedia. Wikimedia Statistics. Disponível em <https://stats.wikimedia.org/#/all-projects>, junho 2021.
- [73] Wikimedia Foundation. Swedish Wikipedia surpasses 1 million articles with aid of article creation bot. Disponível em <https://blog.wikimedia.org/2013/06/17/swedish-wikipedia-1-million-articles/>, junho 2021.
- [74] Wikimedia Foundation. Wikidata:Wikidata Concepts Monitor. Disponível em https://www.wikidata.org/wiki/Wikidata:Wikidata_Concepts_Monitor, junho 2021.
- [75] Wikimedia Foundation. Wikipedia:Administration. Disponível em <https://en.wikipedia.org/wiki/Wikipedia:Administration>, junho 2021.
- [76] Wikipedia. Language proposal policy. Disponível em https://meta.wikimedia.org/wiki/Language_proposal_policy, junho 2021.
- [77] Wikipedia. Lista de Wikipédias. Disponível em https://pt.wikipedia.org/wiki/Lista_de_Wikip%C3%A9dias, junho 2021.
- [78] Wikipedia. Statistics. Disponível em <https://en.wikipedia.org/wiki/Special:Statistics>, junho 2021.
- [79] Wikipedia. The Healthcare Translation Task Force. Disponível em <https://en.wikipedia.org/wiki/Help:Wikitext>, junho 2021.
- [80] Wikipedia. Wikipedia:Content assessment. Disponível em https://en.wikipedia.org/wiki/Wikipedia:Content_assessment, junho 2021.
- [81] Wikipedia. Wikipedia:Five pillars. Disponível em https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Cinco_pilares, junho 2021.
- [82] Wikipedia. Wikipedia:Manual of Style/Medicine-related articles. Disponível em https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Medicine-related_articles, junho 2021.
- [83] Wikipedia. Wikipedia:Researching with Wikipedia. Disponível em https://en.wikipedia.org/wiki/Wikipedia:Researching_with_Wikipedia, junho 2021.
- [84] Wikipedia. Wikipedia:Size of Wikipedia. Disponível em https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia, junho 2021.
- [85] Wikipedia. Wikipedia:Templates. Disponível em <https://en.wikipedia.org/wiki/Wikipedia:Templates>, junho 2021.
- [86] Wikipedia. Wikipedia:What Wikipedia is not. Disponível em https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not, junho 2021.
- [87] Wikipedia. Wikipedia:Why Wikipedia is so great. Disponível em https://en.wikipedia.org/wiki/Wikipedia:Why_Wikipedia_is_so_great, junho 2021.

- [88] Wikipedia. Wikipedia:Wikipedia is not a dictionary. Disponível em https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_dictionary, junho 2021.
- [89] Wikipedia. Wikipedia:WikiProject Medicine/Assessment. Disponível em https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine, junho 2021.
- [90] Wikipedia. Wikipedia:WikiProject Medicine/Participants. Disponível em https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Participants, junho 2021.
- [91] Wikipedia. Wikipedia:WikiProject Medicine/Popular pages. Disponível em https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Popular_pages, junho 2021.
- [92] Wikipedia. Wikipedia:WikiProject Medicine/Popular pages. Disponível em https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Popular_pages, junho 2021.
- [93] Wikipedia. Wikipedia:wikiproject medicine/popular pages. Disponível em https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Popular_pages, 2021.
- [94] Wikipedia. Wikipedia:WikiProject Medicine/Stats/Number of articles by language 2019. Disponível em https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Stats/Number_of_articles_by_language_2019, junho 2021.
- [95] Wikipedia. WikiProject Medicine. Disponível em https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine, junho 2021.
- [96] Wikipedia. Wikipédia:Critérios de notoriedade. Disponível em https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Crit%C3%A9rios_de_notoriedade, junho 2021.
- [97] Wikipedia. Wikipédia:Política de bloqueio. Disponível em https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Pol%C3%ADtica_de_bloqueio, junho 2021.
- [98] Kewen Wu, Qinghua Zhu, Yuxiang Zhao, e Hua Zheng. Mining the factors affecting the quality of Wikipedia articles. *Proceedings - 2010 International Conference of Information Science and Management Engineering, ISME 2010*, 1(1):343–346, 2010.
- [99] Yanxiang Xu e Tiejian Luo. Measuring article quality in Wikipedia: Lexical clue model. *IEEE Symposium on Web Society*, (19):141–146, 2011.
- [100] Xiaolan Zhu e Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. Em *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, página 288–295, New York, NY, USA, 2000. Association for Computing Machinery.