

# Um framework para a construção automática de perfis de pesquisa com base em modelagem de tópicos semânticos

Pablo Cecilio<sup>1</sup>, Antônio Pereira<sup>1</sup>, Leonardo Rocha<sup>1</sup>, Felipe Viegas<sup>2</sup>

<sup>1</sup>UFSJ - São João del-Rei, MG, Brasil

{cecilio, antoniopereira}@aluno.ufsj.edu.br, lcrocha@ufsj.edu.br

<sup>2</sup>UFMG - Belo Horizonte, MG, Brasil

frviegas@dcc.ufmg.br

**Abstract.** *Recent efforts have focused on identifying multidisciplinary teams and detecting co-Authorship Networks based on exploring topic modeling to identify researchers' expertise. Though promising, none of these efforts perform a real-life evaluation of the quality of the built topics. This paper proposes a framework that allows summarizing articles written by researchers to automatically build research profiles and perform online evaluations. We perform a set of experiments, considering the Lattes repository, contrasting two types of evaluation: (1) an offline in which we exploit a traditional metric (NPMI); and (2) an online where researchers evaluate their own built profiles. We observed that using both together is very important for a comprehensive quality evaluation.*

**Resumo.** *Esforços recentes são focados em identificar equipes multidisciplinares e detectar redes de coautoria explorando estratégias de modelagem de tópicos. Apesar da importância, nenhum desses realiza uma avaliação real da qualidade dos tópicos construídos. Este trabalho propõe um framework que permite resumir artigos escritos por pesquisadores para construir perfis de pesquisa automaticamente e realizar avaliações online sobre esses perfis. Foram considerados experimentos considerando o repositório Lattes, contrastando dois tipos de avaliação: (1) uma offline na qual exploramos uma métrica tradicional (NPMI); e (2) um online onde os pesquisadores avaliam seus perfis. O presente trabalho demonstra a importância em combinar ambas avaliações (offline e online) para medir a qualidade de uma solução.*

## 1. Introdução

A evolução da ciência tem demandado soluções para problemas cada vez mais complexos, exigindo para isso equipes multidisciplinares compostas por pesquisadores com expertises diferenciadas. Nesse contexto, primeiramente é necessário identificar as expertises dos pesquisadores, sendo que as suas publicações científicas são potenciais fontes de informação que podem ser usadas como base para uma associação eficaz e confiável desses pesquisadores e suas áreas de conhecimento [de Siqueira et al. 2020]. Uma tarefa essencial e primordial nesse sentido é criar mecanismos para identificar automaticamente os principais tópicos explorados pelos pesquisadores em suas publicações e, consequentemente, sua expertise. A dificuldade nessa tarefa é o fato da produção científica mundial ter apresentado um crescimento nunca antes visto, tanto no número de artigos publicados quanto no número de pesquisadores envolvidos [Gusenbauer 2019].

O presente trabalho propõem um *framework* que permite filtrar, resumir e analisar artigos escritos por pesquisadores e publicados em diferentes plataformas digitais. O *framework* possui quatro blocos de construção principais: (i) Representação de Dados, (ii) Modelagem de Tópicos, (iii) Entidades de Correlação e (iv) Interface. O objetivo do *framework* é extrair informações relevantes de uma coleção de artigos científicos e fornecer uma interface de visualização que apresente graficamente as informações extraídas.

Para representação dos dados quatro tipos diferentes estratégias foram consideradas: (i) a representação TF-IDF [Salton and Buckley 1988]; (ii) TF-IDF utilizando bigramas; (iii) representação CluWords [Viegas et al. 2019], que atualmente é considerado como *state-of-the-art* para tarefas de modelagem de tópicos [Nunes et al. 2021b]; e (iv) uma proposta do presente trabalho, que combina CluWords com bigramas. Na modelagem de tópicos, foi considerado NMF (*Non-negative Matrix Factorization*) [Lee and Seung 1999] para decompor os tópicos latentes, os quais são exploradas para correlacionar os tópicos e artigos, apresentando as informações de autores dos artigos e suas instituições. Por fim, o *framework* utiliza uma interface visual<sup>1</sup> para resumir todas as informações coletadas, destacando os principais tópicos relacionados a cada artigo, autor e instituição. As duas principais questões de pesquisa (QP) que pretendemos responder neste artigo são: **(QP1)** *O framework consegue resumir efetivamente o perfil de pesquisadores, instituições e artigos?* **(QP2)** *O framework é uma ferramenta útil para apoiar avaliações online de estratégias de modelagem de tópicos, complementando as avaliações offline tradicionais?*

Este trabalho possui uma extensa análise considerando um conjunto de dados composto por artigos publicados por pesquisadores brasileiros de Ciência da Computação. Para selecionar os pesquisadores e seus artigos, tomamos como ponto de partida a Plataforma Lattes. A análise é dividida em duas partes, uma quantitativa e outra qualitativa. A primeira se concentra em avaliar instanciações comparando quatro alternativas de modelagem de tópicos com um experimento *offline* que considera uma métrica tradicional da literatura e um experimento de teste A/B (*online*) no qual grupos de 12 pesquisadores avaliaram distintas instanciações do *framework*, pontuando a qualidade dos tópicos atribuídos a seus artigos, perfis e instituições. No experimento *offline*, as estratégias usando a representação CluWords apresentaram os melhores resultados, com destaque para a representação proposta neste trabalho – CluWords+bigramas. Nos experimentos *online*, os resultados para a versão do *framework* foram 20% superiores do que a versão que a representação CluWords, em termos da avaliação dos pesquisadores. Estes resultados respondem positivamente a **QP1**. Embora ambos os experimentos sugerem que a representação CluWords com bigramas é a versão superior do *framework*, algumas diferenças encontradas na avaliação experimental ressaltam que as avaliações *online* e *offline* são complementares, respondendo positivamente a **QP2**. Na avaliação qualitativa, foi avaliado a usabilidade da proposta usando a metodologia SUS [Kocaballi et al. 2018] com os mesmos 48 pesquisadores, os quais avaliaram como excelente a usabilidade.

*Todo o trabalho (implementações, execuções e avaliações dos resultados) foi realizada pelo aluno Pablo Cecilio, sob a orientação do Prof. Leonardo Rocha. O trabalho faz parte de um projeto maior sobre novas abordagens de modelagem de tópicos e contou com a colaboração do aluno de pós-graduação Felipe Viegas na concepção do ambiente experimental e do aluno Antônio Pereira nas análises de resultados.*

---

<sup>1</sup><https://labpi.ufsj.edu.br/lattestopics/> - username: user-test, password: avaliacao

## 2. O Framework

Nesta seção será apresentado o nosso *framework* cujo objetivo é identificar tópicos de pesquisa abordados pelos artigos, autores e instituições. O *framework* é decomposto em quatro blocos de construção: (i) Representação de Dados, (ii) Decomposição de Modelagem de Tópicos, (iii) Entidades de Correlação e (iv) Interface de Sumarização. O presente trabalho vai além da avaliação tradicional de extrair automaticamente informações relevantes de artigos de pesquisa, em que uma métrica offline é utilizada para medir o desempenho da solução. A Interface de Sumarização permite que a instânciação do *framework* seja avaliada online por meio de um Teste A/B, coletando *feedback* dos próprios pesquisadores.

### 2.1. Representação de Dados

Quatro soluções de representação de dados foram consideradas – (i) TF-IDF, (ii) TF-IDF c/ bigramas e (iii) CluWords – que se mostrou eficaz em tarefas de decomposição de tópicos [Viegas et al. 2019], e (iv) CluWords c/ bigramas – uma nova representação de dados que combina a representação CluWords com bigramas. Como vários termos (palavras) compostos definem tópicos de pesquisa, acreditamos que os bigramas podem melhorar a qualidade dos tópicos. As quatro representações são descritas abaixo:

**TF-IDF:** Representação *Bag-of-words* onde cada termo compõe um vetor de comprimento fixo que explora o TF-IDF tradicional como um recurso de ponderação.

$$TFIDF_{d,t} = TF_{d,t} * \log \left( \frac{|\mathbb{D}|}{1 + d_t} \right) \quad (1)$$

**TF-IDF com bigramas:** A representação TF-IDF anterior adicionando as duas palavras adjacentes dos artigos de pesquisa. Para construir a representação com bigramas, exploramos a função *Phrases* do *gensim*<sup>2</sup>. Para reduzir o tamanho do vocabulário, ignoramos todos os bigramas com  $bigramascore(w_a, w_b) < 0.5$ , onde a função  $count(\cdot)$  retorna a ocorrência na coleção,  $min\_count = 3$  e  $|\mathbb{V}|$  representam o tamanho do vocabulário.

$$bigramascore(w_a, w_b) = \frac{(count(w_a, w_b) - min\_count) * |\mathbb{V}|}{(count(w_a) * count(w_b))} > threshold \quad (2)$$

**CluWords:** É uma representação de dados que explora semelhanças de incorporação de palavras e, principalmente, filtrando possíveis ruídos e ponderando-os adequadamente. A solução CluWords aplica três etapas: (i) *Clustering*, (ii) Filtragem e (iii) Ponderação. *Clustering* explora a estratégia de vizinhança mais próxima para capturar relações semânticas entre palavras por meio de modelos de incorporação. Filtragem, filtra o ruído semântico na representação de vizinhança construída na etapa de *Clustering*. A etapa de Ponderação combina as informações semânticas construídas na primeira etapa com a representação Termo-Frequência (TF) *Bag-of-words*.

**CluWords com bigramas:** Difere da representação das CluWords nas etapas de *Clustering* e Ponderação. Na etapa de *Clustering*, utilizamos os bigramas para construir uma incorporação de palavras, capturando a relação semântica entre as palavras. A etapa de Ponderação combina a informação semântica com a representação em bigramas, explorando a informação Termo-Frequência.

---

<sup>2</sup><https://radimrehurek.com/gensim/models/phrases.html>

## 2.2. Decomposição de Modelagem de Tópicos

A etapa de modelagem de tópicos é baseada na estratégia de *Non-negative Matrix Factorization* (NMF) [Lee and Seung 1999], o qual produz uma decomposição “baseada em partes” de relacionamentos latentes de uma matriz não negativa  $A \in \mathbb{R}^{n \times m}$ , onde  $n$  é o número de exemplos (ou seja, documentos) e  $m$  o número de *features* (ou seja, termos). O objetivo é encontrar uma aproximação  $k$ -dimensional de  $A$  ( $k \ll m$ ) em termos de fatores não negativos  $H \in \mathbb{R}^{n \times k}$  e  $W \in \mathbb{R}^{k \times m}$ . A matriz  $H$  codifica a relação entre documentos e tópicos, enquanto a matriz  $W$  codifica a relação entre termos e tópicos.

## 2.3. Entidades de Correlação

Dado uma coleção de dados contendo artigos de pesquisa relacionados à área de Ciência da Computação e ao final, as matrizes  $H$  e  $W$  que relacionam tópicos a artigos e tópicos a palavras, respectivamente. Seguindo o exemplo, considere que o artigo  $i - th$  da matriz  $H$  trata principalmente do tópico “Aplicativos Web”, enquanto o artigo  $j - th$  trata de “Segurança da Informação”. Para este exemplo, cada artigo possui um ou mais autores, assim é possível destacar quais tópicos estão mais relacionados aos autores através das relações entre artigos e tópicos encontrados. Da mesma forma, como os autores de um artigo pertencem a instituições de pesquisa, também é possível destacar as instituições por tópicos, considerando a relação entre artigos e tópicos. A estratégia consiste em manipular as matrizes fornecidas pelo NMF que correlacionam tópicos e artigos, introduzindo as informações dos autores dos artigos e suas instituições, como no exemplo a seguir.



Figura 1. Calculando as contribuições dos autores para os tópicos.

Considere as matrizes  $H$  e  $W$  para três tópicos. Primeiro, cada tópico é identificado analisando a matriz  $H$  e descobrindo quais palavras estão mais fortemente associadas a cada tópico. Assumindo o exemplo em que o primeiro tópico está associado principalmente a “Aplicativos Web”, o segundo a “Machine Learning” e o terceiro a “Segurança da Informação”. Analisando a matriz  $W$  que relaciona documentos e tópicos, tomando como exemplo a primeira matriz da Figura 1, esta contém três artigos, onde cada posição apresenta a “relevância” do tema para o documento. Assim, agrupando e somando os valores de tópicos alcançados para artigos que pertencem ao mesmo autor, levando-nos à segunda matriz da Figura 1. Assumindo que os três artigos da primeira matriz pertencem ao primeiro autor da segunda matriz, inferindo a “relevância” de cada tópico para este autor:

- Tópico “Aplicativos Web”:  $(60 + 45 + 40) = 145$
- Tópico “Machine Learning”:  $(80 + 30 + 50) = 160$
- Tópico “Segurança da Informação”:  $(20 + 70 + 40) = 130$

Considerando a matriz  $W$  completa, esse mesmo processo pode ser aplicado a todos os autores. Podemos calcular a distribuição entre os tópicos pelos quais as obras de cada autor se relacionam. Assim, realizando uma normalização nas linhas que representam os autores na segunda matriz da Figura 1, é possível mensurar o impacto da pesquisa de cada autor sobre cada tópico (Terceira matriz da Figura 1).

## 2.4. Interface de Sumarização

O *framework* possui uma interface de visualização que resume de forma simples e intuitiva os tópicos e suas associações com pesquisadores e instituições. Desta forma, todas as análises propostas nessa seção podem ser realizadas de forma rápida e eficiente. A interface apresenta os dados relacionados aos tópicos e suas correlações com artigos, pesquisadores e instituições por meio de um grande conjunto de metáforas visuais em um aplicativo web, esse fornece ferramentas capazes de apresentar o mesmo resultado de diferentes formas, incluindo gráficos, mapas de calor, tabelas e métodos de busca interativos. Este *Web App* está disponível em nosso site: <https://labpi.ufsj.edu.br/lattestopics/><sup>3</sup>

## 3. Avaliação Experimental

### 3.1. Coleção de Dados

A coleção de dados é composta por artigos publicados por pesquisadores brasileiros de Ciência da Computação em revistas internacionais publicadas em inglês. Para selecionar os pesquisadores e seus respectivos artigos, foi considerado como ponto de partida a Plataforma Lattes<sup>4</sup>. Assim, o título dos artigos relacionados a todos os pesquisadores brasileiros com doutorado em Ciência da Computação e áreas correlatas foram coletados do Lattes. Além do título, também foram extraídos a lista de coautores, que foi consolidada por meio de um processo de desambiguação. Através do *Semantic Scholar*, as palavras-chave e resumos foram coletados utilizando os títulos, bem como os autores dos mesmos. A Tabela 1 mostra as características da coleção de dados após aplicado um pré-processamento sobre os dados. Na avaliação experimental, a coleção de dados foi dividida em cinco intervalos temporais: 2021-2016, 2015-2011, 2010-2006, 2005-2001 e 2000-1966. Esta divisão temporal permite o acompanhamento da evolução das linhas de pesquisa dos autores presentes na coleção de dados.

**Tabela 1. Tabela de resumo de estatísticas.**

	2021 - 2016	2015 - 2011	2010 - 2006	2005 - 2001	2000 - 1966	2021 - 1966
#autores	2,677	2,827	1,894	1,145	586	3,746
#artigos (total)	15,522	12,218	6,476	3,313	2,051	39,580
#artigos (média)	5.7982	4.3219	3.4192	2.8934	3.5	10.5659
#artigos (std)	7.25	5.0746	4.0497	3.3325	6.0939	15.0116
#artigos (min)	1	1	1	1	1	1
#artigos (mediana)	3	3	2	2	2	5
#artigos (max)	78	51	43	39	86	199

### 3.2. Análise Quantitativa

Considerando os experimentos quantitativos, a questão de pesquisa **QP1** foi derivada em duas questões: (q1) *Qual das combinações de representação de dados e estratégias de modelagem de tópicos é a melhor para representar o perfil de artigos, pesquisadores e instituições?* (q2) *Para todos esses casos, há uma diferença significativa entre as alternativas?* Primeiro comparamos as quatro soluções de modelagem de tópicos (Representação de Dados + NMF) considerando uma métrica de qualidade de tópico tradicional na literatura [Nikolenko 2016], a *Normalized Pointwise Mutual Information*

<sup>3</sup>username: user-test, password: avaliacao

<sup>4</sup><https://lattes.cnpq.br/>

(NPMI). Para um determinado conjunto ordenado de palavras principais  $W_t = (w_1, \dots, w_N)$  em um tópicos, o NPMI é calculado como:

$$NPMI_t = \sum_{i < j} \frac{\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-\log p(w_i, w_j)} \quad (3)$$

Os resultados da Tabela 2 apresentam uma comparação *offline* de diferentes períodos temporais. Neste caso, 20 tópicos foram considerados, onde cada tópico é composto por 20 palavras. Todos os resultados possuem significância estatística por meio de teste ANOVA de duas vias com confiança de 95%. A Tabela 2 apresenta os melhores resultados alcançados considerando a solução que combina CluWords com bigramas, seguindo da representação CluWords (até então considerado o estado da arte), TF-IDF c/ bigramas e, por fim, o tradicional TF-IDF. Estes resultados reforçam resultados de trabalhos recentes na literatura [Nunes et al. 2021a, Pedro et al. 2021] que demonstram a alta qualidade dos tópicos construídos por soluções que exploram CluWords para representação de dados. Além disso, esses resultados também enfatizam as melhorias obtidas (marcadas em negrito) pelas melhorias propostas neste artigo (CluWords c/ bigramas) em relação à proposta original (até 3%).

**Tabela 2. Qualidade os tópicos em termos de NPMI de cada solução de modelo de tópicos em diferentes períodos temporais.**

Período	CluWords c/ bigramas	CluWords	TF-IDF c/ bigramas	TF-IDF
2021 - 2016	<b>0.9520</b>	0.9251	0.5151	0.5256
2015 - 2011	<b>0.9550</b>	0.9311	0.5228	0.5524
2010 - 2006	<b>0.9502</b>	0.9373	0.4958	0.5112
2005 - 2001	<b>0.9494</b>	0.9439	0.4827	0.4969
2000 - 1966	<b>0.9477</b>	0.9462	0.5025	0.5250

É muito importante ressaltar que trabalhos da literatura focam principalmente (se não apenas) em avaliações *offline*, considerando métricas tradicionais, como o NPMI. O presente trabalho inclui um experimento de teste A/B (*online*), que é uma contribuição essencial ao trabalho. Este experimento só pôde ser realizado devido à disponibilidade da interface de visualização<sup>5</sup>, que é uma das contribuições do *framework* proposto. Mais especificamente, o *framework* foi instanciado com quatro versões completas, considerando a interface de sumarização. Cada versão do *framework* apresenta uma representação de dados (descritas na Seção 2.1) e a modelagem de tópicos NMF. Doze pesquisadores avaliaram cada um versão do *framework* individualmente, pontuando a qualidade dos temas atribuídos a seus artigos, perfis e instituições, totalizando 48 participantes. Estes participantes são de instituições distintas do Brasil. A proporção de pesquisadores foi mantida nos quatro grupos de acordo com características essenciais, como experiência, número total de publicações, instituições, etc. Cada pesquisador avaliou as três seguintes atividades considerando o período de 2016-2021: **Atividade 1:** *Os três principais temas atribuídos ao seu perfil estão associados à sua linha de trabalho como pesquisador?* **Atividade 2:** *Os três principais temas atribuídos ao perfil da sua instituição estão associados à linha de trabalho de seus pesquisadores?* **Atividade 3:** *Escolha um artigo de sua autoria. Os três principais tópicos destacados no artigo estão associados ao tópico predominante por ele abordado?*

O período temporal desta avaliação temporal foi o período de pesquisa mais recente (2021-2016). Os pesquisadores pontuaram cada uma das três atividades descritas consi-

<sup>5</sup><https://labpi.ufsj.edu.br/lattestopics/> - username: user-test, password: avaliacao

derando as seguintes pontuações: (i) completamente associados - 10pts; (ii) parcialmente associado - 5pts; e (iii) não associado - 0pts. Cada resposta tem uma pontuação associada a ela, que foi considerada para avaliar as variantes do *framework* proposto. Para ilustrar estas pontuações, considere que um pesquisador X avaliou a Atividade 1 para a instanciação TF-IDF do *framework* da seguinte forma: Tópico 1: completamente associado; Tópico 2: totalmente associado; e Tópico 3: não associado. Considerando um segundo pesquisador Y, este avaliou a Atividade 1 para o TF-IDF com instanciação de bigramas: Tópico 1: completamente associado; Tópico 2: parcialmente associado; e Tópico 3: não associado. Nesses casos, a pontuação total associada pelo pesquisador X é 20 (10 + 10 + 0) e pelo pesquisador Y é 15 (10 + 5 + 0). Com base nesses dois pesquisadores e considerando apenas a Atividade 1, a melhor variante da ferramenta seria o TF-IDF. Seguindo este processo, a Tabela 3 resume a pontuação associada por todos os pesquisadores (12 para cada variante).

**Tabela 3. Comparação das pontuações alcançadas por cada estratégia considerando a avaliação *online* para o período 2021-2016.**

Atividade	CluWords c/ bigramas	CluWords	TF-IDF c/ bigramas	TF-IDF
Atividade 1	<b>275</b>	200	205	165
Atividade 2	<b>275</b>	255	225	215
Atividade 3	<b>210</b>	180	165	180
Total	<b>760</b>	635	595	560

Os resultados da Tabela 3 mostram que a instanciação do *framework* usando a representação CluWords c/ bigramas apresenta os melhores resultados, considerando todas as atividades (avaliação do perfil de pesquisadores, instituições e artigos). Estes resultados corroboram a avaliação *offline* apresentada na Tabela 2, além de responder às questões *q1* e *q2* levantadas no início desta subseção e, conseqüentemente, **QP1**. A estratégia CluWords c/ bigramas, proposta no presente trabalho pode ser considerada como uma das contribuições do trabalho.

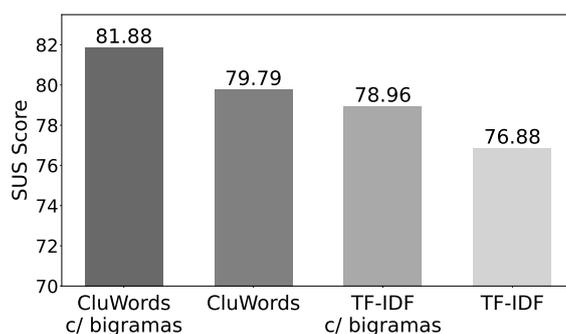
A literatura não apresenta trabalhos que consideram avaliações *online* de estratégias de modelagem de tópicos. A qualidade dos tópicos geralmente é medida explorando apenas métricas tradicionais (experimentos *offline*), como NPMI. Neste contexto, a avaliação *online* permite propor a seguinte questão que estende a **QP2**: (*q3*) – *Existe uma relação entre avaliações offline e online? Quais são as diferenças?* Os resultados apresentados nas Tabelas 2 e 3 foram comparados para responder esta questão. Considerando apenas os resultados referentes ao período 2021-2016, que corresponde ao período avaliado no experimento *online*, pode-se observar alguns contrastes interessantes entre os resultados *offline* e *online*. Do ponto de vista *offline* (NPMI), apesar da diferença estatística de 3% entre as versões do *framework* que consideram as representações CluWords c/ bigramas e o CluWords, ambos apresentam a mesma eficácia. Na avaliação *online*, a avaliação dos pesquisadores apresentaram uma preferência de 20% da versão do *framework* que considera a representação CluWords c/ bigramas em contraste com a versão do mesmo que explora a representação das CluWords (sem bigramas). Este contraste indica que a métrica NPMI captura informações compartilhadas entre as palavras nos tópicos construídos pela instanciação das CluWords. Mas na perspectiva dos usuários, estas palavras não capturam bem a semântica dos tópicos para seus interesses de pesquisa, ou pelo menos, não são superiores quando comparados com versão incluindo bigramas. Além disso, enquanto a avaliação *offline* aponta que a solução TF-IDF é um pouco melhor que a TF-IDF c/ bigramas, os usuários do experimento *online* consideram o contrário.

Embora estas observações não respondam completamente à questão *q3*, elas apontam que as avaliações *online* e *offline* são potencialmente complementares e precisam ser consideradas em conjunto, algo que atualmente é pouco investigado na literatura. *Nesta linha, os experimentos destacam mais uma vez como grande contribuição deste trabalho a interface web que permite aos usuários realizar as avaliações dos tópicos.*

### 3.3. Análise Qualitativa

Esta seção se concentra na avaliação da interface de sumarização. Por meio de experimentação *online*, foi avaliada a usabilidade das instanciações do *framework* proposto, usando a metodologia *System Usabilidade Scale* (SUS) [Kocaballi et al. 2018, Bangor et al. 2008]. A avaliação considerou os mesmos 48 pesquisadores dos experimentos apresentados na seção anterior. O objetivo com este experimento é responder a dois refinamentos de **QP2** (quarta e quinta questões): (*q4*) – *Qual é a avaliação dos pesquisadores sobre a usabilidade de nossa interface de visualização proposta?* (*q5*) – *A qualidade dos tópicos, segundo a avaliação quantitativa, influencia a percepção dos usuários sobre a usabilidade da proposta?* Assim, todos os pesquisadores responderam dez questões, onde cada resposta varia de Concordo totalmente (5pts) a Discordo totalmente (1pts):

1. *Eu acho que gostaria de usar este sistema com frequência.*
2. *Eu acho o sistema desnecessariamente complexo.*
3. *Eu acho o sistema fácil de usar.*
4. *Eu acho que preciso do suporte de um técnico para usar este sistema.*
5. *Eu acho que as várias funções deste sistema estão bem integradas.*
6. *Eu acho que há muita inconsistência neste sistema.*
7. *Eu acho que a maioria das pessoas aprenderia a usar este sistema rapidamente.*
8. *Eu acho o sistema muito complicado de usar.*
9. *Eu me senti muito confiante usando o sistema.*
10. *Eu precisei aprender muitas coisas antes de começar a usar este sistema.*



**Figura 2. Média SUS Score.**

A estratégia utilizada para calcular o *SUS Score*, que representa o grau de usabilidade do sistema, consiste em: Subtrair um em cada resposta das questões ímpares; para as questões pares, a resposta obtida é subtraída por cinco. No final, os valores são somados e multiplicados por 2,5 para converter as pontuações originais de 1-5 para o intervalo de 0-100. A Figura 2 apresenta os resultados referentes a cada versão do *framework*. De acordo com a metodologia do SUS, a pontuação alcançada para todas as versões *framework* foram superiores a 68, o que é considerado uma boa usabilidade, respondendo positivamente à questão *q4*. Observe que a avaliação acima de 80.3 é considerada excelente, que é o caso da instanciação do *framework* que explora a representação CluWords c/ bigramas.

Além disso, analisando as diferenças entre a pontuação do SUS para cada instanciamento e comparando com os resultados apresentados na Tabela 3, existe uma perceptível relação entre os resultados, o que significa que a qualidade dos tópicos construídos influencia diretamente na percepção de usabilidade pelos pesquisadores, respondendo a questão *q5*. A resposta desta última questão fornece evidências de que a usabilidade pode estar intimamente associada à eficácia da solução, mas isso requer uma investigação mais aprofundada em trabalhos futuros.

As dez questões da metodologia do SUS foram analisadas individualmente para responder mais uma questão: (*q6*) – *Quais pontos ainda precisam ser melhorados?* Foram considerados todos os 48 pesquisadores que participaram deste experimento e apenas aqueles que avaliaram a melhor instanciamento do *framework*, comparando ambos os resultados. As respostas para cada questão foram normalizadas pelo maior valor obtido. Os resultados são mostrados na Figura 3. Considerando todos os pesquisadores, observa-se que as questões de pior desempenho foram questões 1 e 9. Ambas estão associadas à eficácia da proposta. Por outro lado, pode-se observar que estas questões não foram um problema para os 12 pesquisadores que avaliaram a melhor instanciamento do *framework*. Estes resultados reforçam as observações de que a qualidade dos tópicos construídos influencia a percepção de usabilidade pelos pesquisadores.

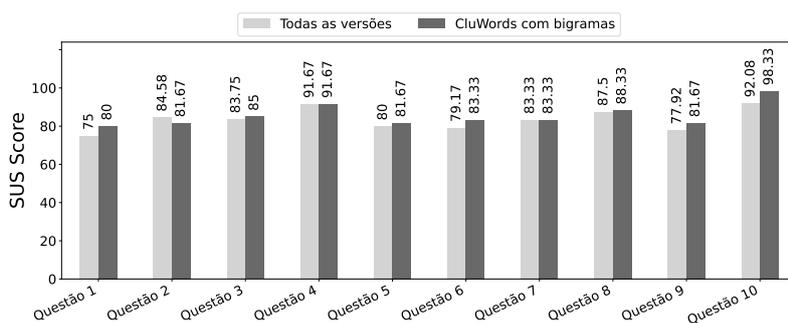


Figura 3. Avaliação individual cada questão do SUS.

#### 4. Conclusão e Trabalhos Futuros

Nesse trabalho foi proposto um *framework* que permite resumir artigos escritos por pesquisadores para construir automaticamente perfis de pesquisa e realizar avaliações *online* quanto à sua qualidade. O *framework* explora e amplia estratégias de modelagem de tópicos de última geração baseadas em CluWords considerando n-grams e apresenta uma nova interface visual capaz de destacar os principais tópicos relacionados a artigos, pesquisadores e instituições.

Para avaliar a capacidade do *framework* em resumir o perfil dessas entidades, bem como sua utilidade para apoiar avaliações *online* da qualidade dos tópicos, foi realizado e contrastado dois tipos de avaliação, considerando um grande repositório de currículos brasileiros: (1) uma avaliação *offline*, na qual foi explorado uma métrica tradicional (NPMI) para medir a qualidade de diversas estratégias de representação de dados, incluindo (i) TF-IDF, (ii) TF-IDF com bigramas, (iii) CluWords e (iv) CluWords com bigramas; e (2) uma avaliação *online* por meio de um teste A/B onde os pesquisadores avaliam seus próprios perfis construídos. Também foi realizado uma avaliação *online* da interface do usuário por meio de um teste de usabilidade seguindo a metodologia SUS. A avaliação experimental

indica que a representação das CluWords com bigramas é a estratégia mais eficaz em avaliações *offline* e *online*. No entanto, ambas as avaliações produziram resultados distintos em termos de proporção da qualidade dos tópicos construídos. A avaliação *offline* apresentou pequenos ganhos de 3% para a melhor instanciação, enquanto a avaliação *online* apresentou grandes melhorias de cerca de 20%. Os resultados apresentam a sensibilidade de construir tópicos em cenários do mundo real. Os resultados de usabilidade da interface apresentam uma relação entre a qualidade da interface e a qualidade dos tópicos percebida pelos pesquisadores, motivando ainda mais as avaliações *online*.

Como trabalho futuro, o *framework* vai ser estendido para novas instanciações explorando em outras estratégias, aprofundando as análises das diferenças e correlações entre avaliações *online* e *offline*, explorando seus pontos fortes e complementaridades.

## Referências

- Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6):574–594.
- de Siqueira, G. O., Canuto, S. D., Gonçalves, M. A., and Laender, A. H. F. (2020). A pragmatic approach to hierarchical categorization of research expertise in the presence of scarce information. *Int. J. Digit. Libr.*, 21(1):61–73.
- Gusenbauer, M. (2019). Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214.
- Kocaballi, A. B., Laranjo, L., and Coiera, E. (2018). Measuring user experience in conversational interfaces: a comparison of six questionnaires. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference*, page 21. BCS Learning & Development Ltd.
- Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.
- Nikolenko, S. I. (2016). Topic quality metrics based on distributed word representations. In *SIGIR'16*.
- Nunes, D., Matos, D., Gomes, J., and Neto, F. (2021a). Chronic pain and language: A topic modelling approach to personal pain descriptions. <https://arxiv.org/abs/2109.00402>.
- Nunes, D. A. P., de Matos, D. M., Ferreira-Gomes, J., and Neto, F. (2021b). Chronic pain and language: A topic modelling approach to personal pain descriptions. *CoRR*, abs/2109.00402.
- Pedro, A., Pereira, A., Cecilio, P., Pena, N., Viegas, F., Tuler, E., Dias, D. R., and Rocha, L. (2021). *An Article-Oriented Framework for Automatic Semantic Analysis of COVID-19 Researches*, pages 172–187. Springer International Publishing.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.
- Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). Cluwords: exploiting semantic word clustering representation for enhanced topic modeling. pages 753–761.