Universidade Federal do ABC

Mensurando o grau de atratividade de textos de divulgação científica usando técnicas de aprendizado de máquina

Projeto de Graduação em Computação I – PGC-I

Aluno
Marcelo de Souza Pena
RA: 11039314
marcelo.pena@aluno.ufabc.edu.br

Orientador
Prof. Dr. Jesús P. Mena-Chalco
jesus.mena@ufabc.edu.br

1 de fevereiro de 2020

Resumo

Diversas pesquisas apontam o alto interesse dos brasileiros por ciência e tecnologia e o crescimento da procura de informações destes temas em mídias sociais como blogs. Técnicas de otimização para motores de busca podem ser utilizadas para aumentar a visibilidade destes blogs, mas para ir além, deve ser possível adequar o conteúdo de forma que ele se torne mais atrativo ao público. Para isso, pretendo utilizar técnicas de aprendizado de máquina na criação de um modelo que permita predizer quantos acessos um texto de divulgação científica terá depois de publicado. A relevância deste projeto recai na possibilidade de auxiliar divulgadores científicos a adequarem seus textos de forma a maximizar o número de visualizações, contribuindo ainda mais para a difusão da ciência.

Palavras-chave: Blogs, divulgação científica, aprendizado de máquina.

Sumário

1 Introdução

O interesse da população por ciência e tecnologia é alto: 61% das pessoas se dizem interessadas ou muito interessadas; no entanto a visão positiva sobre a ciência e sobre os cientistas tem piorado ao longo dos anos, segundo pesquisa[1] do CGEE — Centro de Gestão e Estudos Estratégicos do Ministério da Ciência, Tecnologia, Inovações e Comunicações. A familiaridade da população com o conhecimento científico é precária: 73% dos entrevistados acreditam que antibióticos podem ser usados para matar vírus e 90% não sabiam ou não se lembravam do nome de algum(a) cientista brasileiro(a). Em outra pesquisa[2], feita pelo Datafolha, 43% dos entrevistados discordaram da frase "O ser humano e o chimpanzé vem de uma espécia de origem comum", demonstrando descrença ou desconhecimento sobre a Teoria da Evolução. Dentre os principais meios utilizados na busca de informações sobre ciência e tecnologia estão sites de busca e mídias sociais, tais como *Twitter*, *Instagram*, *YouTube* e *Facebook*. Em outra pesquisa[3], focada em jovens entre 14 e 25 anos, o interesse em ciência também se mostrou alto (67%) e o uso de sites de busca e mídias sociais se fez ainda mais presente.

Dado o interesse da população por ciência e os meios utilizados para se informar sobre o assunto, a divulgação científica em mídias sociais se mostra extremamente relevante.

Dentre os diferentes meios para a divulgação científica, destaco o uso de blogs por estes serem de fácil acesso e serem ótimos meios de difusão do conhecimento. Como exemplos, cito: Science Blogs Brasil[4] que compila conteúdo de uma série de blogs de diversos autores e funciona como um selo que atesta o blog em questão como sendo uma fonte confiável de informações científicas; Blogs Unicamp[5], blog institucional da Universidade Estadual de Campinas que compila conteúdo de diversos blogs feitos por pesquisadores da univer-

[1] "Percepção pública da C&T no Brasil - 2019", disponível em https://www.cgee.org.br/documents/10195/734063/CGEE_resumoexecutivo_Percepcao_pub_CT.pdf

[2] "Vacinas, evolução, transgênicos: pesquisa revela crenças dos brasileiros", disponível em http://revistaquestaodeciencia.com.br/questao-de-fato/2019/05/13/vacinas-evolucao-transgenicos-pesquisa-revela-crencas-dos-brasileiros

[3] "O que os jovens brasileiros pensam da ciência e da tecnologia?", disponível em http://www.coc.fiocruz.br/images/PDF/Resumo%20executivo%20survey%20jovens_FINAL.pdf

[4] Disponível em http://scienceblogs.com.br/, última visita em dezembro de 2019.

[5] Disponível em https://www.blogs.unicamp.br/, última visita em dezembro de 2019.

sidade; e UFABC Divulga Ciência[6], blog institucional da Universidade Federal do ABC que compila conteúdo de diversas iniciativas de divulgação científica na universidade.

Este trabalho tem por objetivo prever o número de acessos que textos em blogs com essa temática terão depois de publicados, possibilitando a adequação do conteúdo a fim de aumentar sua atratividade. Ele está dividido em seis seções além desta introdução.

Pretendo considerar como estudo de caso o blog "Guia dos Entusiastas da Ciência"[7], projeto de extenção da Universidade Federal do ABC do qual faço parte.

Na Seção 2 apresento conceitos teóricos importantes para a total compreensão e contextualização deste trabalho, sendo eles Divulgação científica e sua importância (Subseção 2.1), blogs de divulgação científica (Subseção 2.2) e Search Engine Optimization (Subseção 2.3). Na Seção 3 listo artigos envolvendo aprendizado de máquina e divulgação científica que têm relação com este trabalho. Na seção 4 descrevo os objetivos gerais e específicos que pretendo atingir com este trabalho. Já nas Seções 5 e 6 estão a metodologia a ser empregada, a descrição das fases do projeto e o cronograma a ser seguido nestes doze meses. Finalmente, a Seção 7 discorro sobre a importância e a aplicabilidade do que me proponho a fazer.

2 Conceitos

Nesta Seção fundamento teoricamente conceitos indispensáveis para a total compreenssão deste trabalho.

2.1 Divulgação científica

A divulgação científica pode ser definida como "[...] o trabalho de comunicar ao público, em linguagem acessível, os fatos e os princípios da ciência, dentro de uma filosofia que permita aproveitar o fato jornalisticamente relevante como motivação para explicar os

[6]Disponível em http://proec.ufabc.edu.br/divulgaciencia/, última visita em dezembro de 2019.
[7]Disponível em http://proec.ufabc.edu.br/gec/, última visita em dezembro de 2019.

princípios científicos, os métodos de ação dos cientistas e a evolução das idéias científicas. Aquêle fato jornalisticamente interessante não ocorre todos os dias. Cabe, porém, ao divulgar tornar interessantes os fatos que êle mesmo vai respingando no noticiário. E se tiver habilidade, fará isso até com fatos antigos, que êle trará novamente à vida.", segundo José Reis (Massarani & Alves (2019) apud Reis (1964)), p. 2, precursor da divulgação científica no Brasil, tendo o Prêmio José Reis de Divulgação Científica e Tecnológica sido criado em sua homenagem.

Já para Bueno (2010) apud Bueno (2009), p. 2, ela seria a "[...] utilização de recursos, técnicas, processos e produtos (veículos ou canais) para a veiculação de informações científicas, tecnológicas ou associadas a inovações ao público leigo". Difere, portanto, dos conceitos comunicação científica e jornalismo científico, sendo estes, respectivamente, a difusão de conhecimento científico entre especialistas da área (Bueno, 2010) e a veiculação de informações científica para o público leigo por meios de comunicação de massa, tais como jornais, revistas, rádio, TV ou jornalismo *online* (Bueno (2010) apud Bueno (2009)). Desta forma, todo jornalismo científico é divulgação científica, mas a divulgação científica não se restringe ao jornalismo científico, podendo existir nas mais diversas formas, por exemplo, palestras, livros didáticos, histórias em quadrinhos, animações, memes.

2.2 Blogs de divulgação científica

Blogs são *websites* feitos para expor opiniões, pensamentos ou experiências do autor ou autora na internet. Em especial, blogs de divulgação científica costumam focar seu conteúdo em temas científicos e terem por trás um(a) cientista ou um(a) entusiasta da ciência (Flores, 2016).

A vantagem de divulgar ciência por meio de blogs é a facilidade para publicar e difundir o conteúdo, uma vez que existem diversas plataformas gratuitas como *Wordpress*, *Medium*, *Blogspot*, dentre outros, que não exigem conhecimento técnico para a criação do site.

Nesse contexto, Fausto *et al.* (2017) faz uma análise quantitativa da blogosfera ci-

entífica brasileira através de dados estatísticos.

2.3 SEO

O termo *Search Engine Optimization* (SEO) descreve um conjunto de estratégias que visa tornar seu site melhor classificado por buscadores, fazendo com que ele apareça entre os primeiros resultados das buscas. Dificilmente usuários passam da quinta página de resultados quando querem encontrar uma informação, por isso é importante se atentar a isso (Yalçın & Köse, 2010).

Dentre essas estratégias, destacam-se: uso de palavra-chave, meta-descrição, *links* internos, *links* externos, uso de redes sociais, frequência de atualizações, acessibilidade e usabilidade (Ledford, 2015; Zhang & Cabage, 2013).

O uso dessas técnicas em blogs escolares feitos por alunos aumentou o número de visitantes em 35,3% e o número de comentários em 9,4% comparados a blogs que não usaram SEO (Zhang & Cabage, 2013).

Com o uso de mídias sociais sendo a principal fonte de pesquisa dos brasileiros sobre ciência e tecnologia, o uso de SEO torna-se primordial.

3 Trabalhos correlatos

Blogs de divulgação científica brasileiros têm sido amplamente estudados, sobretudo na área da comunicação (Fausto *et al.* , 2017; Flores, 2016).

Já na área de aprendizado de máquina, blogs (não necessariamente sobre ciência) têm servido de objeto de estudo para diversas pesquisas, seja na detecção de SPAM (Kolari *et al.* , 2006), categorização (Elgersma & de Rijke, 2006) ou análise de sentimentos (Mishne *et al.* , 2005; Yang *et al.* , 2007). Há ainda a tentativa de estimar o fator de impacto de publicações em redes sociais (Schünke, 2015).

São poucos os trabalhos que tentam estimar a atratividade do conteúdo dos blogs,

dentre os quais se destacam os trabalhos de Buza (2014) e Das *et al.* (2016), onde uma série de modelos de regressão foram usados para prever o *feedback* que novos textos de assuntos variados receberiam depois de determinado tempo.

4 Objetivos

Nesta Seção listo os objetivos geral e específicos que pretendo atingir com este trabalho.

4.1 Objetivo geral

Mensurar a atratividade de textos de divulgação científica por meio da aplicação de técnicas de aprendizado de máquina utilizando como base de dados os textos disponíveis em blogs de Ciência. Pretendo considerar como estudo de caso o blog Guia dos Entusiastas da Ciência.

4.2 Objetivos específicos

Dentre os objetivos específicos para esta esta pesquisa considero:

- Criação de uma base de dados de atributos de textos de divulgação científica; e
- Criação de um modelo com alto desempenho na predição do número de visualizações que um texto de divulgação científica terá depois de publicado.

5 Método

Para o desenvolvimento deste projeto pretendo seguir cinco procedimentos (ver Figura 1).

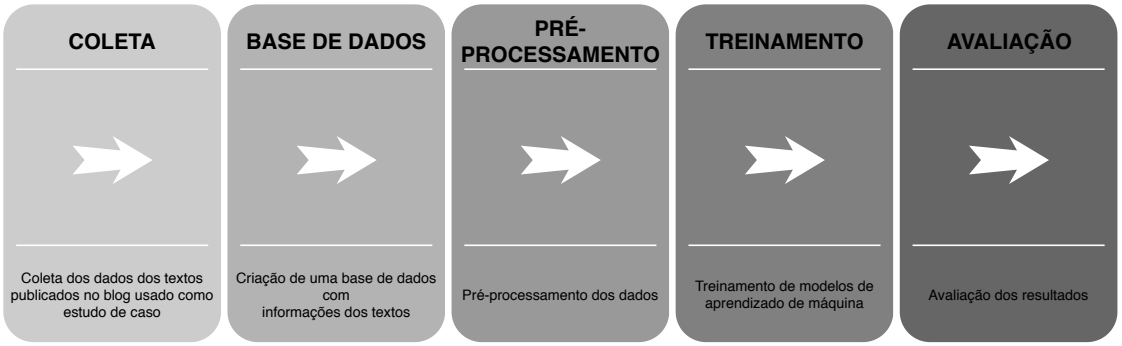

Figura 1: *Passos a serem empregados neste PGC. Cada caixa representa uma tarefa a ser realizada.*

5.1 Coleta

Coletarei os dados do painel de controle de blogs (ainda a ser definido), sendo eles: número de palavras, número de acessos (obtido pelo *plugin WP Statistics*), número de mídias no texto (sejam elas imagens, vídeos ou outros), data de publicação e classificação SEO (obtida pelo *plugin Yoast SEO*), que pode ser dividida em mais itens. Todas as informações ficarão armazenadas em um formato tabular.

5.2 Base de dados

Criarei uma base de dados com as informações supracitadas, além da classificação de complexidade de cada texto (entre 1 e 5, sendo 1 - fácil compreensão e 5 - alta complexidade), grande área do conhecimento em que se enquadra o texto e relevância momentânea do texto na época de sua publicação (a principio, obtida pela ferramenta *Google Trends*).

5.3 Pré-processamento

Nesta etapa farei todo o pré-processamento que se mostrar necessário, como a conversão da data de publicação em dias desde a publicação e posterior aplicação de uma

fórmula que adéque este atributo à dinâmica de conteúdos publicados na internet, onde o pico de visibilidade acontece nos primeiros dias, caindo exponencialmente ao longo do tempo, *e.g.*, não necessariamente um texto com 100 visualizações em uma semana terá 200 visualizações em duas semanas.

5.4 Treinamento

Com o auxílio de ferramentas computacionais sobre Aprendizado de Máquina usarei diferentes técnicas de aprendizado supervisionado para criação de um modelo de regressão para prever o número de acessos de textos de divulgação científica. Pretendo usar, como ferramenta inicial de análise, o software *Weka - Waikato Environment for Knowledge Analysis* (Hall *et al.* , 2009).

5.5 Avaliação

Durante todo o andamento do projeto estudarei técnicas objetivas para avaliação de resultados. Por fim, farei a análise detalhada e avaliação dos resultados obtidos.

6 Cronograma de atividades

O período de realização do projeto é de doze meses e as atividades se darão conforme a Tabela 1. As linhas representam as atividades a serem feitas ao longo dos meses, que são representados pelas colunas. As células em vermelho escuro são as atividades já realizadas, enquanto as em vermelho claro são as que ainda estão por fazer. As três colunas marcadas em azul claro representam os períodos de entrega do PGC I, PGC II e PGC III.

	Atividade	Meses
		1 2 3 4 5 6 7 8 9 10 11 12
1	Levantamento bibliográfico	
2	Coleta dos dados	
3	Criação da base de dados	
4	Pré-processamento dos dados	
5	Treinamento de modelos de aprendizado de máquina	
6	Avaliação e análise dos resultados	
7	Formulação da monografia	
8	Divulgação dos resultados	

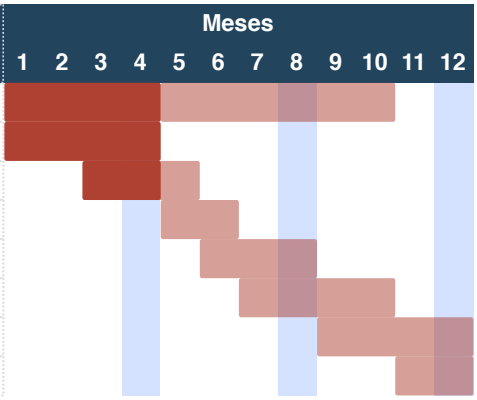

Tabela 1: *Cronograma de atividades e tempo de execução.*

7 Considerações finais

Este projeto permitirá que cientistas, jornalistas e entusiastas da ciência tenham uma referência quando forem produzir textos de divulgação científica, de forma que consigam maximizar seu número de acessos e assim contribuir para a difusão do conhecimento científico.

Devido à pouca produção científica com esta temática, também tenho como objetivo que ele sirva como ponto de partida para futuros trabalhos, de forma que aplicações da inteligência artificial possam auxiliar na popularização da ciência.

Referências

Bueno, Wilson Costa. 2010. Comunicação cientifica e divulgação científica: aproximações e rupturas conceituaiss. *Informação & informação*, **15**(1esp), 1–12.

Bueno, Wilson da Costa. 2009. Jornalismo científico: revisitando o conceito. *Jornalismo científico e desenvolvimento sustentável. são paulo: All print*, 157–78.

Buza, Krisztian. 2014. Feedback prediction for blogs. *Pages 145–152 of: Data analysis, machine learning and knowledge discovery*. Springer.

Das, Shovra, Hasan, Md Hasibul, Rahim, Md Shamsur, & Rahman, Mohammod Hafizur. 2016. A learning dataset aimed at predicting the feedbacks for bengali blogs. *The aiub journal of science and engineering (ajse)*, **15**(1).

Elgersma, Erik, & de Rijke, Maarten. 2006. Learning to recognize blogs: A preliminary exploration. *In: Proceedings of the workshop on NEW TEXT wikis and blogs and other dynamic text sources*.

Fausto, Sibele, Takata, Roberto, Martínez, María Teresa Moreno, Apunike, Alexcolman Tochukwu, Bucci, Jade Lorena Mariano, dos Santos, Ana Carolina Gonçalves, da Silva, Walas João Ribeiro, Matias, Mariane, & Kinouchi, Osame. 2017. O estado da blogosfera científica brasileira. *Em questão*, **23**(5), 274–289.

Flores, Natália Martins. 2016. Entre o protagonismo e a divulgação científica: as estratégias discursivas de constituição do ethos discursivo do cientista blogueiro em blogs de ciência brasileiros. *Tese (tese em ciências da comunicação) - ufpe*.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, & Witten, Ian H. 2009. The WEKA data mining software: an update. *Sigkdd explorations*, **11**(1), 10–18.

Kolari, Pranam, Finin, Tim, Joshi, Anupam, *et al.* . 2006. Svms for the blogosphere: Blog identification and splog detection. *In: Aaai spring symposium on computational approaches to analysing weblogs*.

Ledford, Jerri L. 2015. *Search engine optimization bible*. Vol. 584. John Wiley & Sons.

Massarani, Luisa Medeiros, & Alves, Juliana Passos. 2019. A visão de divulgação científica de josé reis. *Ciência e cultura*, **71**(1), 56–59.

Mishne, Gilad, *et al.* . 2005. Experiments with mood classification in blog posts. *Pages 321–327 of: Proceedings of acm sigir 2005 workshop on stylistic analysis of text for information access*, vol. 19.

Reis, José. 1964. A divulgação da ciência e o ensino. *Ciência e cultura*, **16**(4).

Schünke, Marco Aurélio. 2015. Aplicação de algoritmos de classificação para análise dos fatores que influenciam na predição do fator de impacto nas redes sociais. *Dissetação (dissertação em ciência da computação) - ufrgs*.

Yalçın, Nursel, & Köse, Utku. 2010. What is search engine optimization: Seo? *Procedia-social and behavioral sciences*, **9**, 487–493.

Yang, Changhua, Lin, Kevin Hsin-Yih, & Chen, Hsin-Hsi. 2007. Emotion classification using web blog corpora. *Pages 275–278 of: Ieee/wic/acm international conference on web intelligence (wi'07)*. IEEE.

Zhang, Sonya, & Cabage, Neal. 2013. Does seo matter? increasing classroom blog visibility through search engine optimization. *Pages 1610–1619 of: 2013 46th hawaii international conference on system sciences*. IEEE.