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

4 de junho 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. Resultados preliminares indicam que textos com perguntas no título atraem mais visitantes. Quanto à área, textos de física tem um desempenho muito melhor que textos de história. Como era de se esperar, textos publicados há mais tempo também alcançam um número maior de visualizações. Na predição do número de visualizações o melhor desempenho foi da Regressão Linear, que alcançou um R2 Score de 0.37 depois de aplicada redução de dimensionalidade, normalização e retirada de *outliers*.

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

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.

Com este trabalho pretendo responder a duas perguntas:

- É possível utilizar métodos de regressão para estimar quantas visualizações um texto de divulgação científica terá depois de n dias de sua publicação?
- Existem características que tornam um texto de divulgação científica mais atrativo do que outro? Se sim, quais?

Com base nestas respostas pode ser possível adequar do conteúdo de forma a aumentar sua atratividade.

Como estudo de caso utilizei dados do blog "Guia dos Entusiastas da Ciência"[7], projeto de extensão da Universidade Federal do ABC do qual faço parte desde o início e do qual tenho acesso aos dados necessários para esta análise.

Este trabalho está dividido em seis seções além desta introdução. 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 7 estão a metodologia a ser empregada, a descrição das fases do projeto e o cronograma a ser seguido nestes doze meses. Finalmente, a Seção 8 discorro sobre a importância e a aplicabilidade do que me proponho a fazer.

[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.

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 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 científica 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.* (n.d.), 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 gerais e específicos que busco 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 características dos textos e identificar as características mais relevantes para um maior número de visualizações. Considero 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 com características do textos de divulgação científica

utilizados como estudo de caso;

- Criação de um modelo de predição do número de visualizações que um texto de divulgação científica terá depois de publicado; e
- Criação de um corpus de textos de divulgação científica.

5 Método

Para o desenvolvimento deste projeto segui os seguintes procedimentos (ver Figura 1).

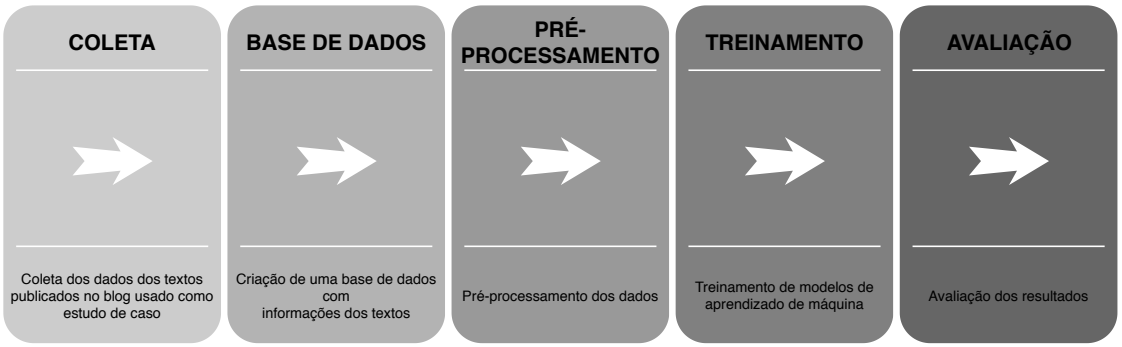

Figura 1: *Passos a serem empregados neste PGC. Cada caixa representa uma tarefa a ser realizada.*

5.1 Coleta

Primeiramente fiz a coleta manual dos textos para a criação do *corpus* de textos de divulgação científica. Os textos coletados foram publicados no blog Guia dos Entusiastas da Ciência[8] no período entre junho de 2018 e fevereiro de 2020. O *corpus* conta com 118 textos, com título, conteúdo e autoria, somando mais de 1600 parágrafos.

Para a base de dados coletei manualmente no painel de controle do blog as seguintes informações de cada texto:

[8]Disponível em `http://proec.ufabc.edu.br/gec/`, última visita em maio de 2020.

- Títulos;
- Data de publicação;
- Usuário pelo qual o texto foi publicado;
- Categoria na qual foi publicado, sendo estas:
  - Ciência ao redor (como a ciência está presente no seu dia a dia);
  - O que que a ciência tem? (explicações sobre termos científicos);
  - Profissão Cientista (vida e carreira de cientistas);
  - Sci... what? (artigos publicados em revistas científicas explicados com linguagem didática e acessível);
  - Ciência Pop (a ciência presente em produtos da cultura pop como filmes e séries);
  - ABC da ciência (textos sobre projetos produzidos na Universidade Federal do ABC);
  - Você disse ciência? (esclarecimentos sobre pseudociências e *Fake News* envolvendo ciência); e
  - Outros.
- Área do conhecimento ao qual o texto se encaixa, sendo estas:
  - Ciência;
  - Química;
  - Biologia;
  - Física;
  - História;
  - Medicina;
  - Astronomia;

  - Matemática;
  - Psicologia;
  - Atualidades; e
  - Tecnologia.

- Quantidade de mídias, sendo elas imagens, vídeos ou outros ('Mídia');
- Classificaçao SEO dada pelo *plugin Yoast SEO* ('SEO'), sendo;
  - 1 para textos classificados com SEO Bom; e
  - 0 para textos classificados com SEO OK ou ruim.
- Número de *links* internos ('Links I.');
- Número de *links* externos ('Links E.');
- Complexidade ('Complexidade') em uma análise subjetiva feita por mim em que:
  - 1 indica que o texto é adequado a todos os públicos, necessitando apenas de uma compreensão do nível de Ensino Básico;
  - 2 indica a necessidade de Ensino Médio; e
  - 3 indica a necessidade de Ensino Superior.
- Outras características subjetivas do texto recomendadas por Massarani *et al.* (2004) e utilizadas aqui como variáveis binárias:
  - Introdução ('Introdução') que indica a presença ou não de um parágrafo introdutório no inicio do texto ao invés de partir logo para o assunto;
    - 1 para textos que possuem introdução; e
    - 0 para textos que não possuem.
  - Analogias ('Analogias') que indica o uso ou não de analogias ao longo do texto para facilitar explicações;
    - 1 para textos que possuem analogias; e

    * 0 para textos que não possuem.

  - Interação ('Interação') que indica se o texto utiliza ou não linguagem impessoal que conta com a participação do leitor; e
    * 1 para textos que possuem interação; e
    * 0 para textos que não possuem.
  - Siglas ('Siglas') que indica se o texto possui ou não uma descrição específica do significado significado de toda sigla que utiliza, sem presumir que o leitor a conheça. Obs: algumas siglas mais conhecidas pela sigla do que por seu significado não foram consideradas nesta avaliação, tais como NASA ou GPS.
    * 1 para textos que possuem siglas com descrição específica; e
    * 0 para textos que não possuem.

- *Link* direto para o texto no blog; e
- Número de visualizações ('Visualizações'), fornecido pelo *plugin WP Statistics*.

Considerei a utilização de ferramentas como o *Google Trends* para estimar um valor relativo à procura por palavras-chave de cada texto, mas isso se mostrou inviável por dois motivos: as principais ferramentas fazem análises comparativas entre os termos buscados, de forma que não é possível saber números absolutos, apenas períodos em que os termos foram mais ou menos buscados com um intervalo de 0 a 100, sendo 100 o momento de mais buscas. O segundo motivo é o fato de que o número de visualizações coletado diz respeito às visualizações em todos os dias entre a data de publicação do texto e a data da coleta, independente do assunto do texto ter tido maior ou menor visibilidade em parte desse período. Uma análise dia-a-dia comparando número de buscas e número de visualizações para comprovar a correlação e estudar a causalidade entre elas poderia ser feita, mas foge do escopo deste projeto.

5.2 Base de dados

Para a criação da base de dados, além das informações coletadas manualmente, coletei um conjunto de características por meio das técnicas de Processamento de Linguagem Natural *tokenização* e *postagging* das bibliotecas *NLTK* e *Spacy* e algoritmos na linguagem Python utilizando o *corpus* com os textos. Essas características são:

- Uso ou não de perguntas como título ('Pergunta');
- Número de palavras ('numPal');
- Número de parágrafos ('numPar');
- Número de substantivos ('numSub');
- Número de adjetivos ('numAdj');
- Número de verbos ('numVrb');
- Número de Entidades Nomeadas ('numNEs');
- Número de determinantes ('numDet');
- Número de conjunções ('numConj');
- Número de advérbios ('numAdv');
- Número de adposições ('numAdp');
- Número de numerais ('numNum');
- Número médio de caracteres nos parágrafos ('tamParagraf'); e
- Número de caracteres no título ('tamTítulo').

Todas as informações ficaram armazenadas em uma planilha.

5.3 Pré-processamento

Para tornar alguns dados utilizáveis, foram aplicadas as seguintes transformações:

- A data de publicação foi trocada para dias desde a publicação até o último dia de fevereiro de 2020 (data da coleta das visualizações);
- As categorias e áreas, que eram dados categóricos, foram transformadas em variáveis binárias através do método *get_dummies* da biblioteca *pandas*, assim cada uma delas se tornou uma coluna com o nome "categoria-Nome da Categoria"ou "área-Nome da Área";
- Quatro *outliers* foram retirados da base de dados, sendo estes textos com mais de 2000 visualizações sem causa parente, como mostra a Figura 2;
- Uma segunda base de dados foi criada através do *MinMaxScalerScaller* da biblioteca *sklearn*, que transforma os valores máximo e mínimo encontrados em determinada variável em 1 e 0, respectivamente, com os demais valores assumindo valores intermediários proporcionais; e
- Uma terceira base de dados foi criada através do *StandardScaller* da biblioteca *sklearn*, que tenta aproximar cada variável de uma curva normal, com média zero e desvio padrão um.

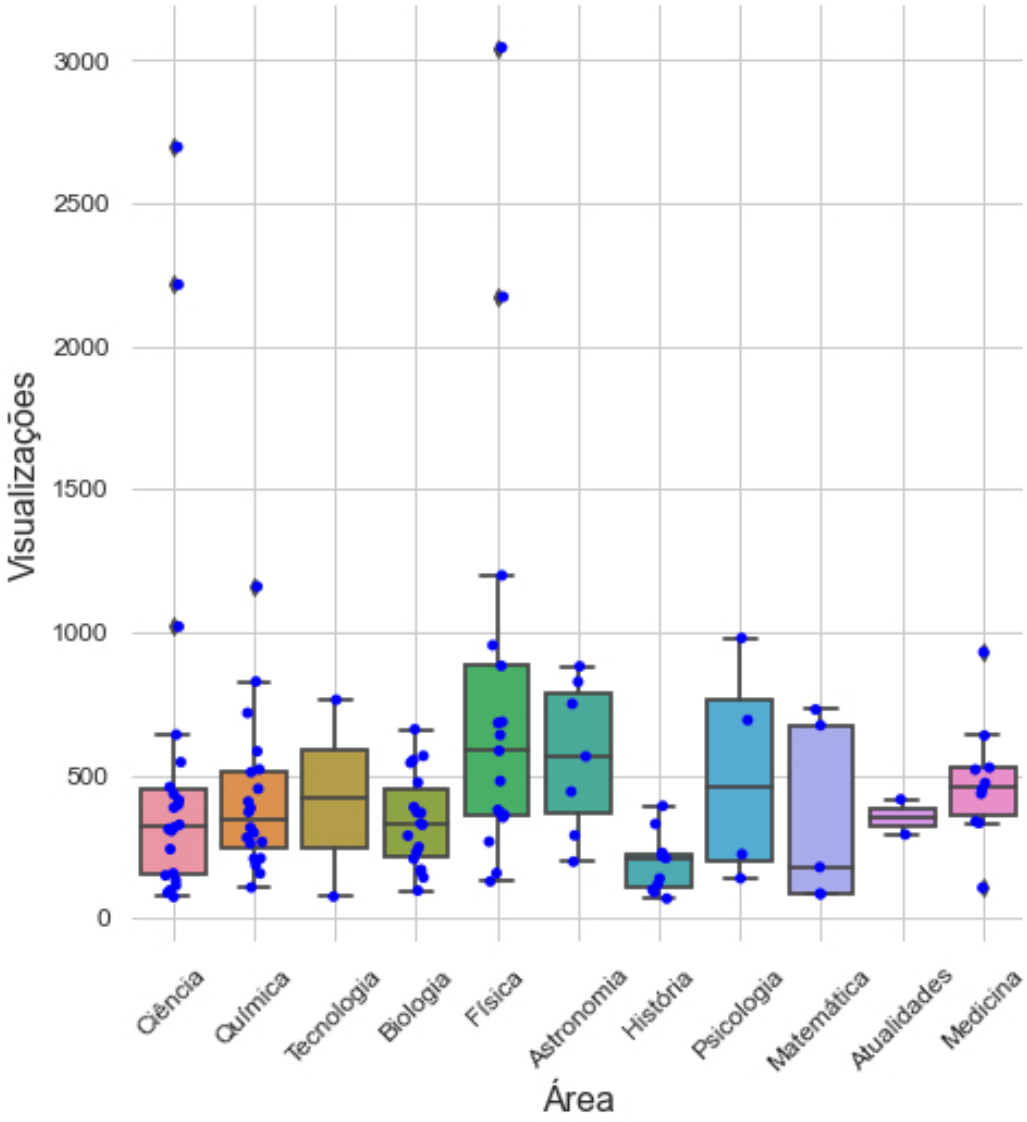

Figura 2: *Visualização de* outliers.

Um exemplo pode ser visto na Figura 3, onde constam os valores originais e transformados do primeiro texto da base.

	DF Completo	MinMax Scaller	Standard Scaller
Mídia	1	0.000000	-0.502990
SEO	0	0.000000	-2.309401
Links I.	0	0.000000	-0.352154
Links E.	2	0.036364	-0.622700
Complexidade	1	0.000000	-1.304985
Introdução	1	1.000000	0.342997
Analogias	0	0.000000	-0.404061
Interação	1	1.000000	0.570597
Siglas	1	1.000000	0.502740
numPal	345	0.091377	-1.071791
numPar	9	0.075000	-0.621779
numSub	137	0.123552	-0.799781
numAdj	30	0.144144	-0.876618
numVrb	41	0.064639	-1.137234
numNEs	47	0.165414	-0.022993
numDet	39	0.079545	-1.203229
numConj	11	0.097222	-0.945167
numAdv	15	0.074380	-1.090766
numAdp	49	0.098361	-1.076549
numNum	4	0.057143	-0.579258
Pergunta	1	1.000000	1.261312
tamParagraf	308	0.222042	-0.844538
tamTitulo	40	0.392857	-0.065875
Dias	634	1.000000	1.502015
categoria-ABC da ciência	1	1.000000	5.244044
categoria-Ciência Pop	0	0.000000	-0.214176
categoria-Ciência ao redor	0	0.000000	-0.749429
categoria-O que que a ciência tem?	0	0.000000	-0.679366
categoria-Outros	0	0.000000	-0.235702
categoria-Profissão Cientista	0	0.000000	-0.358766
categoria-Sci... what?	0	0.000000	-0.255774
categoria-Você disse ciência?	0	0.000000	-0.133631
área-Astronomia	0	0.000000	-0.255774
área-Atualidades	0	0.000000	-0.133631
área-Biologia	0	0.000000	-0.433013
área-Ciência	1	1.000000	2.167948
área-Física	0	0.000000	-0.389249
área-História	0	0.000000	-0.326797
área-Matemática	0	0.000000	-0.214176
área-Medicina	0	0.000000	-0.310087
área-Psicologia	0	0.000000	-0.190693
área-Química	0	0.000000	-0.461266
área-Tecnologia	0	0.000000	-0.133631

Figura 3: *Valores para o primeiro texto da base de dados.*

A Figura 4 mostra a correlação entre os atributos. A correlação entre todas as quantidades de palavras é alta, enquanto as maiores correlações com o número de visualizações são 0,546 com 'Dias' e 0,395 com 'Pergunta'.

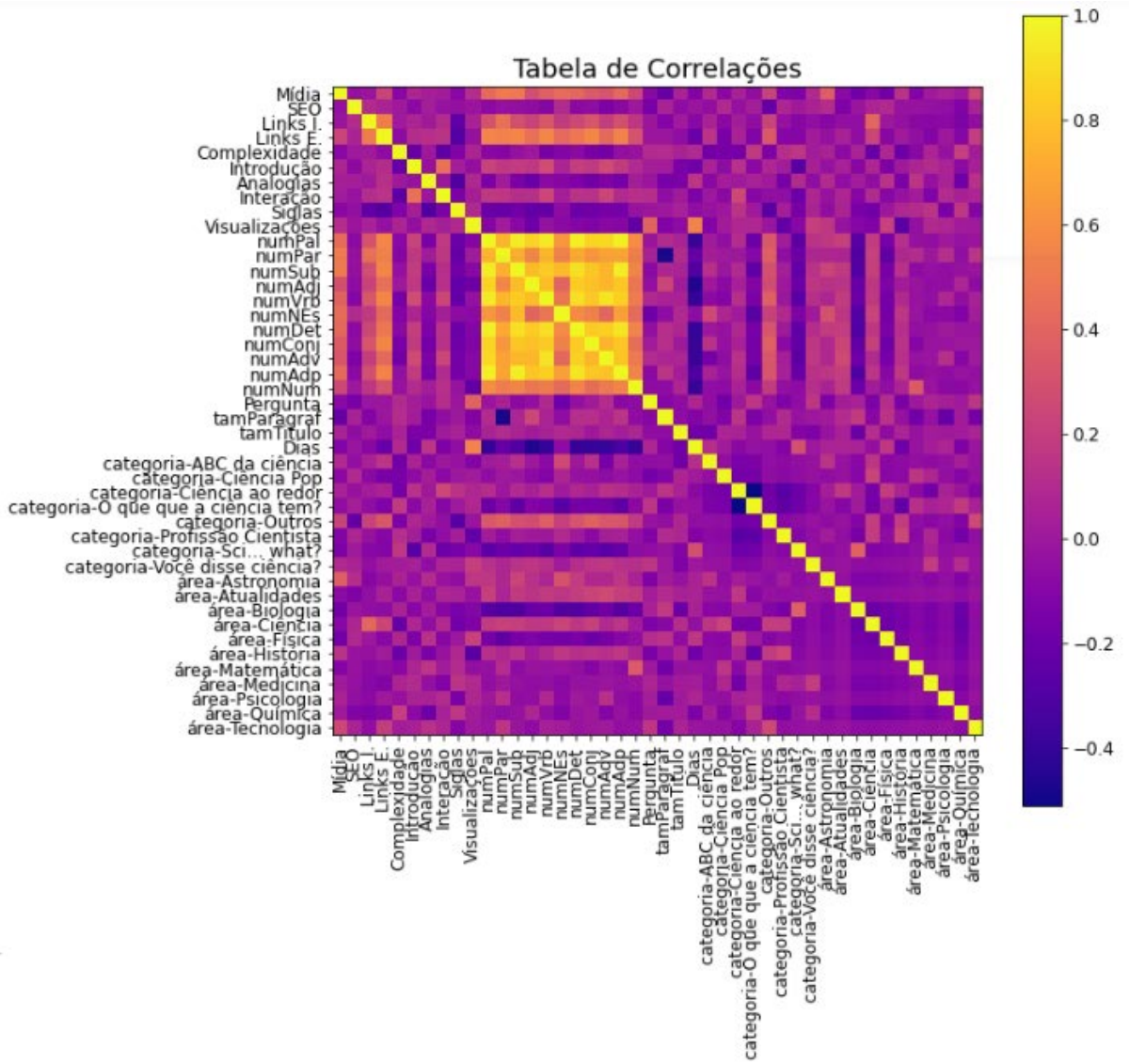

Figura 4: *Tabela de correlação.*

A distribuição dos atributos binários pode ser vista na Figura 5.

	0	0 (%)	1	1 (%)
categoria-ABC da ciência	110	96.5	4	3.5
categoria-Ciência Pop	109	95.6	5	4.4
categoria-Ciência ao redor	73	64.0	41	36.0
categoria-O que que a ciência tem?	78	68.4	36	31.6
categoria-Outros	108	94.7	6	5.3
categoria-Profissão Cientista	101	88.6	13	11.4
categoria-Sci... what?	107	93.9	7	6.1
categoria-Você disse ciência?	112	98.2	2	1.8
área-Astronomia	107	93.9	7	6.1
área-Atualidades	112	98.2	2	1.8
área-Biologia	96	84.2	18	15.8
área-Ciência	94	82.5	20	17.5
área-Física	99	86.8	15	13.2
área-História	103	90.4	11	9.6
área-Matemática	109	95.6	5	4.4
área-Medicina	104	91.2	10	8.8
área-Psicologia	110	96.5	4	3.5
área-Química	94	82.5	20	17.5
área-Tecnologia	112	98.2	2	1.8
Introdução	12	10.5	102	89.5
Analogias	98	86.0	16	14.0
Interação	28	24.6	86	75.4
Siglas	23	20.2	91	79.8

Figura 5: *Distribuição dos atributos binários.*

5.4 Treinamento

Para o treinamento as características *Título*, *Links* e *Usuário* foram retiradas da base. As duas primeiras só foram coletadas para que fosse possível identificar os textos e o nome de usuário do autor foi desconsiderado pois as peculiaridades de cada autor já devem ser consideradas nas características coletadas com Processamento de Linguagem Natural, como o tamanho médio dos parágrafos e se o título é uma pergunta ou não, sendo assim utilizar o nome do usuário seria redundante.

Na etapa de treinamento apliquei os seguintes métodos utilizando os melhores parâmetros avaliados pelo método *Grid Search* e com todas as reduções de dimensionalidade entre 2 e o total de colunas menos 1 utilizando o método *PCA*:

- *Random Forest Regressor* com número de árvores 10, 50, 100, 150, 200, 250 e 300 e número mínimo de amostras nos nós folha de 1, 5, 10, 15 e 20;
- *SVM Regressor com kernel linear* com epsilon de 0.05, 0.1, 0.15, 0.2 e 0.25;
- *Linear Regression* com e sem normalização;
- *MLP Regressor* com 10, 20, 50, 100, 150, 200, 250 e 300 camadas; e
- *KNeighbors Regressor* com 1, 3, 5, 15, 25, 35, 45 e 55 vizinhos, utilizando peso uniforme ou de distância e com os algoritmos auto, balltree, kdtree e brute.

Também considerei utilizar SVM Regressor com kernel RBF e com kernel Sigmoid, porém eles não se adequaram ao tipo de dados por terem resultados bem abaixo dos demais e portanto foram retirados da análise.

5.5 Avaliação

A avaliação de cada modelo foi feita pelo erro quadrático médio e pelo *R2 Score*.

6 Resultados preliminares

Dentre os resultados mais relevantes, destaco na Figura 6 a linha de tendência e o intervalo de confiança dos textos cujos títulos são perguntas (laranja) e do textos textos cujos títulos não são (azul).

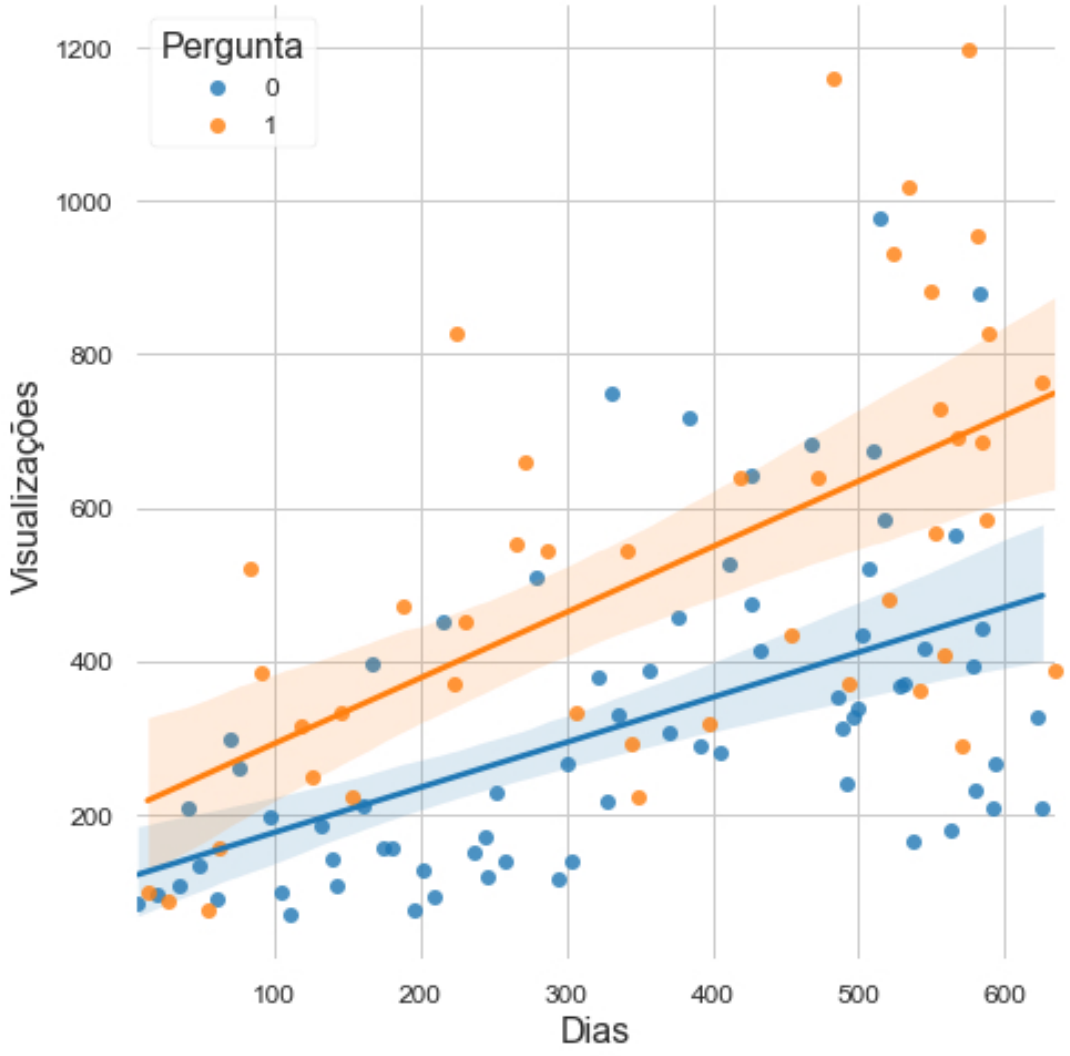

Figura 6: *Visualizações x Dias*

Na Figura 7 mostro o *R2 Score* de cada um dos 5 métodos para o os dados originais com entre 2 e 43 atributos. A Regressão Linear tem o melhor desempenho chegando a um *R2 Score* de 0.37 com PCA 20.

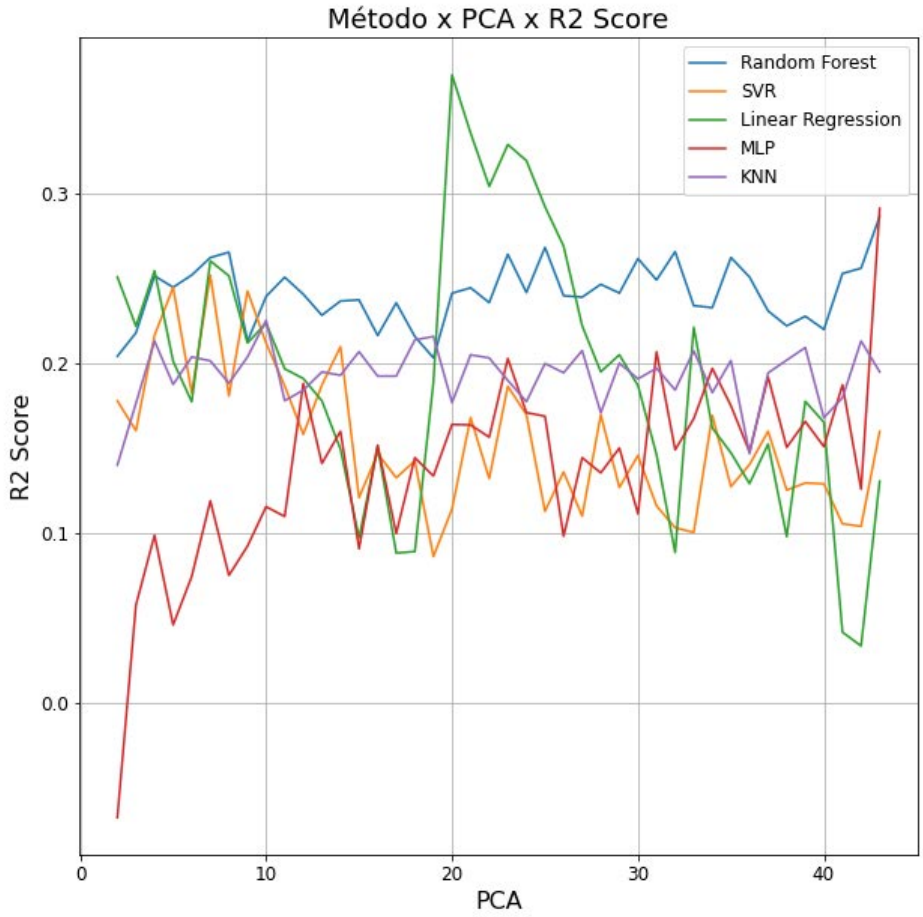

Figura 7: *Comparação*

7 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.*

8 Considerações finais

Com este projeto viso fornecer a cientistas, jornalistas e entusiastas da ciência formas de aumentar o número de visualizações quando forem produzir textos de divulgação científica, de forma que consigam assim contribuir ainda mais 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.

Pelos resultados preliminares, textos com perguntas no título parecem se sair melhor. Quanto às áreas dos textos, aparentemente física desperta muito mais interesse do que história, por exemplo. Como era de se esperar, textos publicados há mais tempo também alcançam um número maior de visualizações.

Quanto à predição do número de visualizações, a tarefa se provou nada trivial. Técnicas de redução de dimensionalidade aliadas à normalização dos dados e à retirada de *outliers* fizeram com que uma Regressão Linear alcançasse um *R2 Score* de 0.37, o mais alto de todos os métodos. Espero que novos testes ajudem a melhorar ainda mais este valor.

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. A learning dataset aimed at predicting the feedbacks for bengali blogs.

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.

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, *et al.* . 2004. Guia de divulgação científica. *Rio de janeiro: Scidev. net: Brasília, df: Secretaria de ciência e tecnologia para a inclusão social*.

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.

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.