INTELIGÊNCI

NA ANÁLISE DE TEXTOS

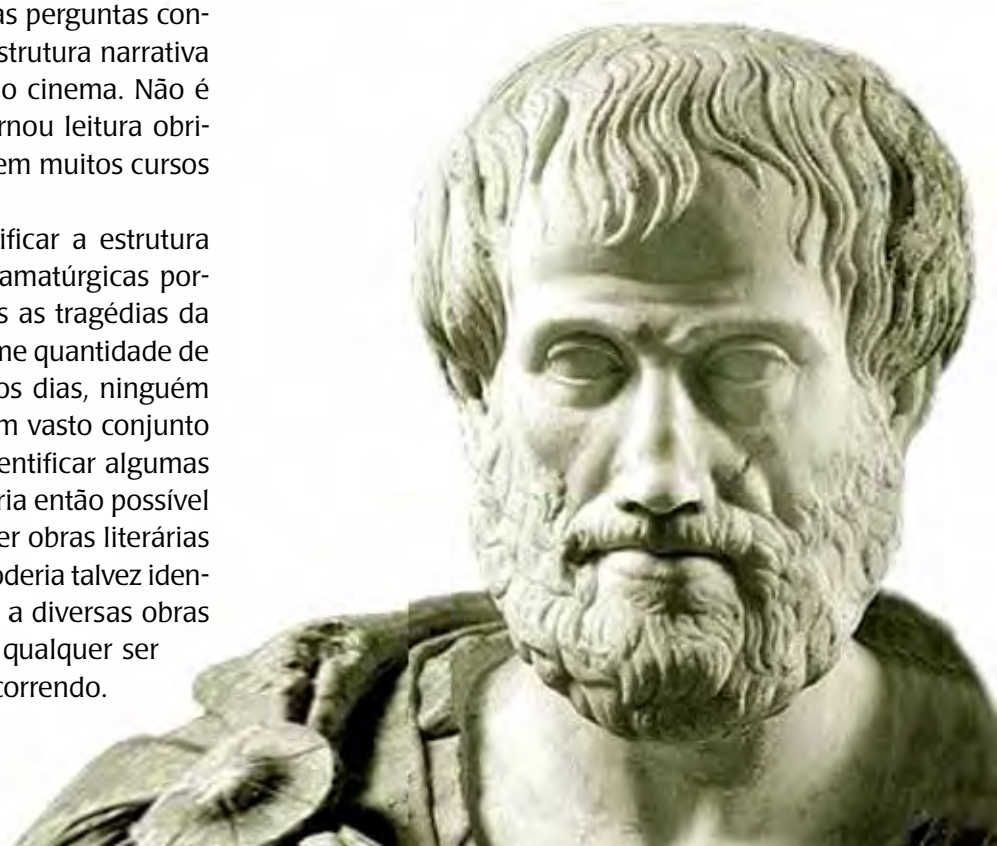
Prof. Dr. Marcelo de Araujo*

Aristóteles escreveu na Antiguidade um texto conhecido como *Poética*, ainda hoje considerado um clássico da teoria literária. Na obra, Aristóteles trata de examinar a estrutura típica de grandes peças de teatro. Quais são os elementos constitutivos de uma boa tragédia? Qual é a estrutura típica de uma narrativa trágica bem-sucedida? A resposta que Aristóteles deu a essas perguntas continua exercendo influência sobre a estrutura narrativa de muitos romances e roteiros para o cinema. Não é por acaso, aliás, que a *Poética* se tornou leitura obrigatória entre roteiristas e é adotada em muitos cursos de escrita criativa.[1]

Aristóteles só foi capaz de identificar a estrutura narrativa típica de grandes obras dramatúrgicas porque ele conhecia praticamente todas as tragédias da Antiguidade. No entanto, face à enorme quantidade de obras de ficção publicadas em nossos dias, ninguém mais pode ter a expectativa de ler um vasto conjunto de obras literárias na tentativa de identificar algumas estruturas narrativas comuns. Não seria então possível delegarmos a máquinas a tarefa de ler obras literárias em nosso lugar? Uma máquina não poderia talvez identificar os "arcos emocionais" comuns a diversas obras literárias com mais precisão do que qualquer ser humano? Na verdade, isso já vem ocorrendo.

Medindo arcos emocionais

Em 2016, Andrew Reagan e colegas publicaram um artigo intitulado "Os arcos emocionais das histórias são dominados por seis formas básicas".[2] Um algoritmo desenvolvido pelos pesquisadores, batizado de "Hedonometer", analisou 1.327 obras literárias, disponíveis no site do Projeto Gutenberg. Cada obra foi dividida



A ARTIFICIAL LITERÁRIOS

em "janelas" ou segmentos de 10 mil palavras. Cada janela foi submetida então a uma "análise de sentimentos". A análise consiste na avaliação quantitativa dos sentimentos que as palavras, que ocorrem nas janelas, despertam no leitor. Palavras como, por exemplo, "assassino" e "roubo" tendem a provocar nos leitores uma reação negativa (uma atitude de reprovação), por oposição a palavras como "honestidade" ou "vitória".

O Hedonometer criou então um dicionário que contém as 10 mil palavras mais frequentes no conjunto das obras analisadas. A cada palavra do dicionário foi atribuído um valor que varia entre 1 e 9. Os valores foram atribuídos graças ao trabalho de milhares de pessoas especialmente recrutadas para essa tarefa. Palavras que têm uma conotação negativa receberam um valor baixo, por oposição às palavras que têm uma conotação positiva. O valor 5 (intermediário entre 1 e 9) indica que a palavra é emocionalmente neutra, não desperta nenhum sentimento especial no leitor. A palavra "carbono", por exemplo, é geralmente neutra; preposições, da mesma forma, também não despertam nenhum tipo de associação emocional no leitor. As três palavras que receberam a maior pontuação média foram, respectivamente, "riso", "felicidade" e "amor". As três últimas palavras no ranking foram "estupro", "suicídio" e "terrorista".[3]

A ocorrência dessas palavras, em cada segmento de 10 mil palavras, permite ao Hedonometer avaliar quantitativamente a carga emocional predominante em cada segmento da obra, e retraçar as flutuações emotivas ao longo da obra como um todo. São essas flutuações emotivas que Reagan e colegas denominam de "arco emocional" da narrativa.[4] A análise de sentimento realizada pelo Hedonometer consiste na representação gráfica das flutuações emotivas ao longo de cada obra analisada. Segundo Reagan e colegas, é possível detectar, no conjunto das 1.327 obras analisadas, seis tipos básicos de arcos emocionais.

Uma história com final feliz, por exemplo, é marcada por um arco ascendente na parte final, diferentemente de narrativas com finais trágicos, que são marcadas por um arco emocional descendente. O artigo de Reagan e colegas, porém, não é o único trabalho recente que descreve o modo como algoritmos podem ser utilizados para ler grandes quantidades de textos literários com o objetivo de analisar certas estruturas comuns, inerentes a praticamente qualquer obra de ficção.

Detectando best-sellers

Em 2016, Jodie Archer e Matthew Jockers lançaram um livro chamado *O Segredo do Best-seller*. A dupla desenvolveu um programa chamado "Bestseller-ometer" na expectativa de poder identificar potenciais best-sellers. O programa leu mais de 20 mil romances com o objetivo de identificar características típicas dos livros que entram para a lista dos mais vendidos do *New York Times*. A descrição técnica do programa aparece no último capítulo do livro de Archer e Jockers. Mas o que me interessa aqui não é examinar a descrição técnica do algoritmo, mas sim chamar atenção para algumas

implicações que a difusão de programas como o Hedonometer e o "Bestseller-ometer" poderiam ter para o mercado editorial e para a nossa compreensão do conceito de "leitor".

O número de manuscritos que editoras e agências literárias recebem todos os dias costuma ultrapassar bastante a capacidade que seus funcionários têm para ler e avaliar todo esse material. Histórias de livros que se tornaram sucessos literários, mas que foram inicialmente rejeitados ou simplesmente ignorados por várias editoras, se tornaram famosas. Mas isso geralmente ocorre, não porque os autores rejeitados sejam gênios incompreendidos, mas porque os profissionais do mercado literário muitas vezes não conseguem dar conta do volume de leitura que recebem. Muitas editoras e agências literárias têm de contratar leitores externos, que decidem então quais manuscritos merecem ser avaliados para possível publicação.

Segundo Archer e Jockers, o "Bestseller-ometer" teria 80% de chance de detectar um manuscrito que tem o potencial para se tornar um autêntico best-seller. Se algoritmos desse tipo se tornarem correntes no mercado editorial, então, no futuro, os primeiros "leitores" de muitas obras de ficção não serão mais seres humanos, mas máquinas que, para todos os efeitos, estarão realizando o mesmo tipo de atividade que os leitores contratados por editoras e agências literárias realizam. Novos escritores, ávidos para publicar seu primeiro romance, talvez prefiram então buscar o aval de algoritmos ao invés de consultar escritores experientes ou críticos literários. É difícil prever de que modo isso poderia interferir no processo de criação literária de escritores e escritoras no futuro.

Por outro lado, é possível também que muitos romances, que têm o potencial para se tornar um sucesso literário, sejam rejeitados com menos frequência, pois haverá um novo "leitor", mais rápido e eficiente, atuando no mercado. Além disso, algoritmos como o Hedonometer e o "Bestseller-ometer" poderiam traçar um painel da produção literária de um dado país, em uma dada época, ou em uma língua específica, e encontrar aí tendências de que nem os escritores nem os profissionais do mercado editorial estão inteiramente conscientes. Conhecer melhor essas tendências é importante, inclusive, para o próprio trabalho de escritores e escritoras – e não só por razões comerciais.

Considere, por exemplo, a pesquisa pioneira de Regina Dalcastagnè, da Universidade de Brasília. Dalcastagnè e sua equipe leram e analisaram centenas de romances brasileiros publicados em três períodos distintos: de 1965 a 1979, de 1990 a 2004, e de 2005 a 2014.[5] Os dados – para mencionar aqui apenas os do segundo período – são reveladores sobre quem são os personagens que habitam os romances publicados por autores brasileiros. A pesquisa analisou o perfil de 1.245 personagens que aparecem em 258 romances de escritores e escritoras brasileiros publicados entre 1990 e 2004, e constatou o seguinte:

79,8% são brancos;

56,6% são da classe média;

81% são heterossexuais; e

71,1% dos protagonistas são homens.

Entre os personagens negros, 20,4% são bandidos e 12,2% são empregadas (ou empregados) domésticas. Além disso, a maior parte dos personagens vive no eixo Rio-São Paulo. É comum às vezes pensarmos que, no Brasil, a propaganda, as capas de revistas em bancas de jornal, as telenovelas e outros meios de representação da

sociedade brasileira reproduzem e consolidam estereótipos acerca de pessoas negras, mulheres e homossexuais. Mas a literatura, nesse quesito, aparentemente não é muito diferente. Os protagonistas dos romances brasileiros são também os "protagonistas" da vida real – aqueles que tomam decisões nos tribunais, geram empresas ou criam leis. Numa entrevista sobre sua pesquisa, Dalcastagnè declara de modo bastante informal a sua opinião sobre a literatura brasileira contemporânea: "É tudo muito repetitivo, os enredos, as preocupações, as cidades; muito pouco variado, sem graça. Por que temos tão poucos protagonistas cabeleireiros, manicures, bancários, motoristas de ônibus?".[6]

Algoritmos como o Hedonometer e o "Bestsellerometer", evidentemente, não substituem o trabalho de pesquisadores como Dalcastagnè, mas eles podem se tornar importantes aliados nesse tipo de pesquisa. Afinal, a pesquisa de Dalcastagnè se limitou à análise de romances publicados por três grandes editoras no

Brasil: a Record, a Companhia das Letras e a Rocco. A produção literária publicada sob a forma de contos não foi levada em consideração. Ficaram também de fora da pesquisa romances policiais e livros de ficção científica.[7] Romances publicados por pequenas editoras, ou autopublicados em plataformas como as da *Amazon* e *Wattpad*, também foram desconsiderados. Mas não poderia estar surgindo hoje – graças em parte à emergência de novas tecnologias no mercado editorial – um novo perfil de personagem na ficção brasileira? Os recursos necessários para investigar essa questão, utilizando-se o mesmo tipo de metodologia empregada por Dalcastagnè, teriam de ser bem mais elevados; a pesquisa exigiria também, com certeza, a formação de uma equipe mais numerosa. A utilização de algoritmos para a análise estatística de um número bastante elevado de textos de ficção poderia auxiliar nesse tipo de tarefa. Um trabalho bastante ambicioso, dessa natureza, foi realizado em 2018 por uma equipe de pesquisadores nos Estados Unidos.

Ted Underwood e colegas utilizaram um algoritmo para ler 104.000 obras de ficção publicadas em inglês entre 1703 e 2009.[8] O algoritmo foi capaz de identificar o sexo dos personagens com base nos nomes atribuídos em 90% dos casos. Contra o que se poderia talvez esperar, o algoritmo constatou que o espaço dispensado à caracterização de personagens do sexo feminino – medido em número de palavras – não aumentou, mas diminuiu com o passar dos anos. A esse fenômeno os autores deram o nome de "masculinização da ficção". Foi apenas a partir da década de 1960 que se pode constatar um gradual (e discreto) aumento do espaço destinado à caracterização de personagens do sexo feminino. Para a identificação do sexo do personagem o algoritmo recorreu não apenas aos nomes próprios utilizados, mas também ao tipo de linguagem e ao vocabulário empregado ao se caracterizar os personagens. Aqui, a equipe constatou outro resultado interessante. Nos romances escritos no século XIX, esse método de identificação do sexo do personagem era confiável em 75% dos casos. Ou seja: a linguagem e o modo de caracterização do personagem variavam conforme o sexo do personagem, e isso permitia ao algoritmo determinar se o personagem em questão era do sexo feminino ou masculino. A utilização de palavras como, por exemplo, "coração", "lágrimas", "suspiros" e "sorriso" estavam mais associadas à caracterização de personagens do sexo feminino do que do sexo masculino. Mas a partir do século XXI

esse método de identificação se tornou menos confiável: o algoritmo, com base nesse método, acertou em apenas 65% dos casos. Os autores sugerem que isso ocorreu porque, com o tempo, a distinção entre a linguagem e o vocabulário tipicamente utilizados na caracterização de personagens masculinos ou femininos foi se diluindo.

A utilização de algoritmos nos departamentos de literatura deve provavelmente se tornar cada vez mais frequente daqui para a frente. Já existem, inclusive, diversas ferramentas especialmente desenvolvidas para pesquisadores e estudantes da área de literatura.[9] Esses novos métodos de pesquisa nos obrigam a repensar e ampliar o conceito de "leitor" no âmbito da produção literária contemporânea.

Lendo e aprendendo

Essa ampliação do conceito de "leitor", porém, tem implicações jurídicas também. Em setembro de 2016, pesquisadores da *Google* publicaram um artigo no qual descrevem o funcionamento de um algoritmo desenvolvido para gerar frases em linguagem natural.[10] O algoritmo leu mais de 11 mil obras de ficção para que as frases geradas por ele fossem estilisticamente melhores do que as frases geradas por outros algoritmos, que também são capazes de gerar textos em linguagem natural.

Empresas como *Google* e *Facebook* vêm investindo bastante na criação de *chatbots* ou assistentes virtuais, capazes de responder a perguntas de usuários e de manter uma conversa coerente sob a forma de *chats* on-line. Programas desse tipo, na verdade, não são nenhuma novidade. Joseph Weizenbaum já tinha criado um (ELIZA) na década de 1960. O problema, porém, é que programas como ELIZA contam com um estoque limitado de frases prontas, que são reutilizadas com alguns ajustes gramaticais conforme o *input* do interlocutor. Isso torna a interação com o programa monótona e pouco natural. Para evitar esse problema, a *Google* e outras empresas pretendem desenvolver agora assistentes virtuais inteligentes, capazes de gerar frases novas, que soem menos artificiais e que não sejam diretamente extraídas de um banco de frases prontas. Para isso, é necessário que o assistente virtual leia milhares de obras com o objetivo de identificar uma diversidade de padrões e estilos de conversação, mas sem repetir literalmente as frases que lê.

O artigo publicado pelos pesquisadores da *Google*, no entanto, gerou um problema jurídico. As obras de ficção lidas pelo algoritmo não estavam em domínio público. No momento em que essas obras foram disponibilizadas on-line, ninguém havia ainda considerado a possibilidade de que, entre os seus leitores, estariam também algoritmos, capazes de ler milhares de obras e de reutilizá-las para fins comerciais. Muitos escritores e escritoras se sentiram lesados ao saberem que suas obras haviam sido lidas por algoritmos, e não por seres humanos. Pela declaração que deram à imprensa após a divulgação do caso, é possível perceber que, para todos os efeitos, os autores e autoras dos textos veem os algoritmos como leitores, sujeitos às mesmas restrições jurídicas a que os leitores humanos estão também submetidos. Em uma reportagem do *The Guardian* sobre o ocorrido, alguns dos escritores e escritoras, cujas obras foram usadas na pesquisa da *Google*, deram declarações como essas:

"Talvez eu esteja pensando de modo antiquado, que o leitor lerá meu livro – nunca havia me ocorrido que uma máquina poderia ler o meu livro. [...]" e "A pesquisa

em questão usa esses romances para o exato propósito de seus autores – para serem lidos. [...]". [11]

O uso de algoritmos para a análise de obras de ficção não se limita à "leitura" de romances de maior apelo comercial. O uso se estende também à análise de clássicos da literatura. Pesquisadores poloneses desenvolveram em 2016 um algoritmo para analisar textos de autores como, por exemplo, James Joyce, Virginia Woolf e Roberto Bolaño. Os pesquisadores constataram que muitos clássicos da literatura, diferentemente de best-sellers, têm uma estrutura fractal. Isso significa dizer que o tamanho das frases, contado em número de palavras, vai se alternando segundo padrões específicos. Esses padrões conferem à narrativa um ritmo próprio, uma cadência da qual os leitores (e talvez até mesmo os autores) nem sempre estão inteiramente conscientes.[12]

No contexto da Antiguidade, Aristóteles ainda estava em condição de conhecer praticamente todas as obras dramáticas relevantes e de examinar certas estruturas comuns a todas elas. Nos dias de hoje, porém, nenhum ser humano consegue ter, sozinho, essa visão de todo. A inteligência artificial, eu acredito, não substituirá o trabalho de críticos literários.

Mas, ainda assim, pode muito bem se tornar uma ferramenta indispensável para a análise da estrutura narrativa de obras literárias no futuro. ■

Texto extraído do livro *Novas Tecnologias e Dilemas Morais*, disponível em nossa biblioteca.

Notas

[1] Tierno, Michael. (2002). Aristotle's Poetics for screenwriters: Storytelling secrets from the greatest mind in Western civilization. New York: Hachette Book. Hiltunen, Ari. (2001). Aristotle in Hollywood: The anatomy of successful storytelling. Bristol: Intellect.

[2] Reagan, A. J.; Mitchell, L.; Kiley, D. et. al. (2016). "The emotional arcs of stories are dominated by six basic shapes". EPJ Data Science, vol. 5, n. 31, p. 1-12.

[3] Hedonometer: http://hedonometer.org/words.html.

[4] Arco emocional das obras analisadas pelo Hedonometer: http://hedonometer.org/books/v3/31/.

[5] Dalcastagnè, Regina. (2012). Literatura brasileira contemporânea: Um território contestado. Rio de Janeiro: EdUERJ. Ver também Massuela, Amanda. (2018). "Quem é e sobre o que escreve o autor brasileiro" (entrevista com Regina Dalcastagnè). Revista Cult, 5 de fevereiro de 2018.

[6] Massuela, Amanda. (2018). "Quem é e sobre o que escreve o autor brasileiro" (entrevista com Regina Dalcastagnè). Revista Cult, 5 de fevereiro de 2018.

[7] Dalcastagnè, Regina. (2012). Literatura brasileira contemporânea: Um território contestado. Rio de Janeiro: EdUERJ, p. 151.

[8] Underwood, Ted; Bamman, David; Lee, Sabrina. (2018). "The transformation of gender in English-language fiction". Journal of Cultural Analytics, 25p. (doi: 10.31235/osf.io/fr9bk).

[9] Ver por exemplo Michel, Jean-Baptiste; Shen, Yuan Kui; Aiden, Aviva Presser et alia. (2011). "Quantitative analysis of culture using millions of digitized books". Science, vol. 331, p. 176-182. Jockers, Matthew. (2014). Text analysis with R for students of literature. Springer: Heidelberg e New York. Jockers, Matthew. (2013). Macroanalysis. Digital methods and literary history. Champaign (Illinois): University of Illinois Press. Ver também site do Laboratório de Literatura da Universidade de Stanford (Stanford Literary Lab): https://litlab.stanford.edu/.

[10] Bowman, S. R.; Vilnis, L.; Vinyals, O. et alia. (2016). "Generating sentences from a continuous space". Cornell University Library: https://arxiv.org/abs/1511.06349.

[11] Lea, Richard. (2016). "Google swallows 11,000 novels to improve AI's conversation. As writers learn that tech giant has processed their work without permission, the Authors Guild condemns 'blatantly commercial use of expressive authorship'", 28 de setembro de 2016.

[12] Drożdż, S.; Oświęcimka, P.; Kulig, A. et alia. (2016). "Quantifying origin and character of long-range correlations in narrative texts". Information Sciences, vol. 331, p. 32-44.

* *Doutor em Filosofia pela Universidade de Konstanz, Alemanha. Professor de Filosofia do Direito na UFRJ e Professor de Ética na UERJ.*