Notas de investigación
Research notes
Notas de pesquisa

methaodos.revista de ciencias sociales (2023) 11(2) m231102n03

Produção de texto automático no jornalismo desportivo português: estudo exploratório do Prosebot/Zerozero.pt

Automatic text in Portuguese sports journalism: exploratory research on Prosebot/Zerozero.pt

João Canavilhas | jc@ubi.pt | Autor de correspondência
Universidade da Beira Interior, Portugal

Adriana Gonçalves | adriana.goncalves@ubi.pt
Universidade da Beira Interior, Portugal

10.17502/mrcs.v11i2.682

Recebido: 02-03-2023
Aceito: 02-06-2023

Resumo

Uma das tendências no campo das novas tecnologias aplicadas ao jornalismo é o uso de ferramentas de Inteligência Artificial. As experiências não são novas, mas nos últimos anos ganharam importância por serem uma das opções para ultrapassar a crise que afeta o jornalismo. Na linha de alguns trabalhos que comparam textos escritos por jornalistas e por algoritmos, este estudo exploratório analisa as diferenças entre os textos produzidos pelo algoritmo Prosebot do jornal desportivo português Zerozero.pt e os textos produzidos por jornalistas de várias publicações portuguesas. Com essa finalidade, optou-se por realizar uma análise de conteúdo utilizando a ferramenta Voyant, tendo sido analisadas variáveis como o número e variedade das palavras, o número de parágrafos e palavras por frase, a densidade do vocabulário e a legibilidade dos textos. Além da análise linguística, foi ainda realizada uma avaliação relacionada com o uso das normas jornalísticas. As conclusões permitem dizer que os conteúdos escritos pelo Prosebot respeitam o uso das normas jornalísticas e não apresentam erros de ortografia nem de sintaxe que obriguem a uma revisão humana antes da publicação. No entanto, os textos automáticos são mais curtos, mais pobres em termos linguísticos e seguem uma estrutura muito semelhante, o que os torna menos apelativos. Nesse sentido, a produção automática de texto emerge como uma boa ferramenta de apoio ao jornalismo, mas, para já, ainda não deve ser considerada uma ferramenta com autonomia para produzir notícias.

Palavras-chave: automação, Inteligência Artificial, desporto, futebol, jornalismo automatizado.

Abstract

One of the trends in the field of new technologies applied to journalism is the use of Artificial Intelligence tools. The experiments are not new, but in recent years they have increased importance for being one of the options to overcome the crisis affecting journalism. Aligned with some works that compare texts written by journalists and by algorithms, this exploratory reserach analyses the differences between the texts produced by the Prosebot algorithm of the Portuguese sports newspaper Zerozero.pt and the texts produced by journalists from several Portuguese publications. For this purpose, it was chosen to perform a content analysis using Voyant tools, having been analysed variables such as the number and variety of words, the number of paragraphs and words per sentence, the vocabulary density, and the readability of the texts. Besides the linguistic analysis, an evaluation related to the use of journalistic standards was also performed. The conclusions allow us to say that the contents written by Prosebot respect the use of journalistic standards and have no spelling or syntax errors that require human review before publication. However, the automatic texts are shorter, poorer in linguistics and always follow a very similar structure, which makes them less appealing. In this sense, the automatic production of text emerges as an excellent support tool for journalism, but, for now, it can still be considered as an autonomous way of producing news.

Keywords: automation, Artificial Intelligence, sport, soccer, automated journalism.

Sumário

Como citar este artigo

Canavilhas, J. e Gonçalves, A. (2023). Produção de texto automático no jornalismo desportivo português: estudo exploratório do Prosebot/Zerozero.pt. methaodos.revista de ciencias sociales, *11*(2), m231102n03. http://dx.doi.org/10.17502/mrcs.v11i2.682

1. Introdução

O surgimento dos media online, o sucesso das redes sociais e a recessão económica global, colocaram os modelos tradicionais do negócio mediático em causa, obrigando as empresas a efetuarem despedimentos de profissionais. O jornalismo mergulhou numa crise que não é cíclica, como algumas anteriores, mas sim estrutural (De Mateo *et al*., 2010), e está relacionada com as mudanças num ecossistema mediático em que o aparecimento de novos meios alterou as lógicas de mercado.

Foi assim com o aparecimento da rádio e, mais tarde, com a televisão, mas em nenhuma destas situações os impactos foram tão duradouros como no caso da Web. Este meio trouxe novos concorrentes e modelos de negócio ainda não totalmente compreendidos pelas empresas mediáticas, desencadeando o aparecimento de fenómenos e tecnologias que têm como força motriz o online. É o caso dos smartphones, que introduziram novas variáveis no ecossistema e colocaram mais desafios ao jornalismo.

O surgimento de novas tecnologias com impacto no jornalismo acelerou e, em pouco tempo, a atividade viu-se confrontada com a necessidade de integrar nas suas rotinas tecnologias tão diferentes como os drones, a realidade virtual, o *big data* ou a Inteligência Artificial (Perez-Seijo e Vicente, 2022). Algumas destas tecnologias permitiram melhorar o processo produtivo e colocaram o jornalismo na quarta revolução industrial (Schwab, 2016) ao substituir a automatização de tarefas rotineiras por máquinas com capacidade para raciocinar e aprender por si mesmas.

O recurso a ferramentas de Inteligência Artificial, tema deste trabalho, está na base desta quarta revolução industrial e tem como objetivo equilibrar a perda de recursos humanos com o recurso a sistemas automáticos inteligentes. Desde a recolha de informações (Diakopoulos, 2019), até à produção e à distribuição de conteúdos (Helberger, 2019), são muitos os campos em que a automação se tornou numa alternativa, por isso a Inteligência Artificial transformou-se num precioso auxiliar para o jornalismo em crise (Túñes-López, 2021).

Uma das áreas em grande desenvolvimento é a produção automática de notícias, algo que se justifica por ser uma forma de manter um intenso fluxo de notícias sem necessidade de recursos humanos. Este processo de automatização começou por ocorrer em temas onde existia abundância de dados estruturados, como o desporto ou a economia, mas alargou-se a outros temas com iguais características, existindo, atualmente, exemplos deste tipo de produção um pouco por todo o mundo.

Este trabalho centra-se num caso português –o Prosebot– um algoritmo desenvolvido pela Faculdade de Engenharia do Porto e pelo jornal desportivo online Zerozero.pt para escrever sínteses de jogos de futebol. O sistema acede a uma base de dados interna com as fichas de cada jogo e produz conteúdo textual, recorrendo a técnicas de processamento de linguagem natural (NLP), um subcampo da Inteligência Artificial.

As sínteses automáticas elaboradas pelo Prosebot funcionam como resumos instantâneos dos jogos, que ficam disponíveis no site sem necessidade de passarem por um processo de supervisão humana. Considerando o número de jogos que decorrem todos os fins-de-semana em campeonatos nacionais e distritais nos vários escalões etários, o Prosebot apresenta um enorme potencial em termos de escalabilidade, permitindo a cobertura de jogos que não teriam qualquer exposição mediática por falta de jornalistas para o fazer.

Apesar das potenciais vantagens do sistema, existem igualmente alguns riscos associados à Inteligência Artificial, uns relacionados com questões de ordem ética (Ventura-Pociño, 2022), outros com o papel dos jornalistas neste novo paradigma (Carreira e Squirra, 2017) e, por fim, problemas relacionados com a fiabilidade e a qualidade do texto automático, algumas variáveis que serão analisadas neste trabalho.

Embora seja uma temática lateral aos interesses desta investigação, devem ainda ser mencionados os riscos relacionados com o uso da Inteligência Artificial (IA) na produção de informação falsa, não de forma propositada com recurso a bots (Pena, 2019), mas de forma voluntária, como resultado do mau tratamento automático de dados. O funcionamento autónomo de algoritmos nas redações, e a opacidade em volta dos critérios que levaram ao seu desenvolvimento, poderá criar um ambiente propício à produção e difusão de informação tendenciosa e, por isso, desinformativa, mas com o selo de qualidade dos meios de comunicação social. Sabendo que nas redes sociais circula muita informação falsa produzida por usuários (Chadwick e Vaccari, 2019), os leitores costumam ter algum cuidado no seu processamento, mas se essa informação mal contextualizada e pouco contrastada tiver origem nos media, a tendência é para que os leitores a aceitem como boa. Nesta situação, os processos de desinformação tenderão a aumentar devido à má utilização da IA no jornalismo, um risco a ter em conta e que deve exigir legislação sobre transparência nos algoritmos.

2. Produção automática de conteúdos

Uma das alterações introduzidas pela Internet na vida social foi o aumento da quantidade de informação em circulação. Esta mudança levou os media a criarem versões online para aproveitarem a informação que não cabia nas suas edições periódicas, procurando dessa forma concorrer com as novas plataformas que começavam a surgir, nomeadamente os blogues e as redes sociais.

Numa segunda fase, o ativismo digital e as novas leis da transparência obrigaram as empresas e os poderes públicos a justificarem as suas ações através da publicação aberta de dados relativos à atividade, gerando-se assim um imenso mar de informação. Estes dados transformaram-se num dos pilares da 4ª Revolução Industrial por alimentarem os sistemas algorítmicos, sendo a matéria-prima para a oferta de produtos ou serviços. No caso do jornalismo, a chamada Inteligência Artificial Generativa passou a transformar estes dados em textos, gráficos, sons, imagens, vídeos, etc.

No campo do texto, que é o tema deste trabalho, a produção automática assume várias denominações, como jornalismo algorítmico (Dörr, 2016), jornalismo automático (Carlson, 2015; Graefe, 2016) ou jornalismo robot (Kim e Kim, 2018), referindo-se aos sistemas que transformam dados estruturados em texto escrito. Este processo desenvolve-se a partir da técnica de geração de linguagem natural (NLG), assente num planeamento de três fases. Num primeiro momento, é determinado o que se procura comunicar e como se devem estruturar essas informações em parágrafos e frases. Seguidamente é definida a ordem das informações, de acordo com as prioridades editoriais. Por fim é delineado um microplaneamento, que consiste nas escolhas a efetuar ao nível da sintaxe, e que terão impacto na variabilidade e complexidade da linguagem (Diakopoulos, 2019).

Estes sistemas de automação de texto resultam da interseção entre o jornalismo e o big data (Carlson, 2015), reunindo profissionais das duas áreas. O desenvolvimento destes sistemas tem sido melhorado pelo trabalho conjunto entre empresas tecnológicas e grupos de media, mas Torrijos (2021) nota que o impacto destas ferramentas se reflete mais no aumento dos lucros e da produtividade do que na melhoria direta dos textos. Apesar disso, neste ponto de desenvolvimento há casos em que já é difícil distinguir os textos escritos por algoritmos, daqueles que são produzidos por jornalistas (Edwards *et al.*, 2014), tendo sido efetuados alguns estudos neste campo. Num trabalho realizado por Clerwall (2014), os participantes avaliaram o texto automático como mais descritivo, mais informativo, mais aborrecido, mas também mais preciso, confiável e objetivo, em relação ao texto escrito pelos jornalistas. Já o estudo de Kieslich *et al.* (2021), que avaliou as perceções do público sobre notícias automáticas, revelou uma importante contradição: embora os inquiridos vejam pouca ou nenhuma melhoria nos conteúdos produzidos pela Inteligência Artificial, dizem que, em algumas áreas, o algoritmo tem um melhor desempenho do que os humanos. Esta perceção pode ser explicada pelo facto de o texto automático incluir mais dados estatísticos (Murcia Verdú *et al.*, 2022), algo que tem grande importância em notícias sobre desporto ou economia, os temas onde o texto automático é mais usado.

A maior vantagem da produção automática de textos é a libertação dos jornalistas de tarefas rotineiras, deixando-os mais disponíveis para a criação de conteúdos diferenciados, como reportagens em profundidade (Diakopoulos, 2019). Por isso, a produção algorítmica de notícias é uma das transformações mais significativas nas indústrias de media (Napoli, 2014), desafiando os limites da profissão de jornalista e as formas tradicionais de criar e difundir notícias (Tandoc *et al.*, 2022). Esta situação prende-se com o reposicionamento do papel do jornalista em todo o processo, reforçando as funções de verificação, interpretação e atribuição de sentido às informações (Quandt *et al.*, 2021) em detrimento da atividade de redação. De certa forma, a introdução destas tecnologias obriga os jornalistas a recuperarem a essência da sua profissão (Peña-Fernández *et al.*, 2023), ou seja, a reforçarem a aplicação dos princípios definidos nos manuais da profissão e nos códigos deontológicos, em vez de serem meros relatores de acontecimentos, que por chegarem ao conhecimento do público através de plataformas não jornalísticas, acabam por ter menos importância para os utilizadores.

Para além deste reposicionamento profissional, a IA acelera ainda a adaptação dos jornalistas ao trabalho colaborativo com os novos profissionais na produção noticiosa, como os programadores, os designers e os analistas de dados, que são conhecidos como tecnoatores (Canavilhas *et al.*, 2014).

Deve ainda ser referido que a automatização poderá influenciar as relações organizacionais: Beckett (2019) alerta para a ambiguidade na relação entre os grupos de media e as empresas tecnológicas, desde logo, pela diferença de modelo de negócio e de cultura organizacional. Neste sentido, nem sempre as prioridades jornalísticas se alinham com os interesses financeiros das tecnológicas, porque as empresas tecnológicas detêm o capital financeiro e os media pretendem manter os seus valores tradicionais (Beckett, 2019). Para

contornar estes desafios, alguns media optaram por desenvolver os seus próprios algoritmos, evitando as tensões com as tecnológicas especializadas nesta área.

3. Breve história da produção automática

No campo do jornalismo, a produção automática começou nas especialidades onde existiam grandes volumes de dados disponíveis, como a economia e o desporto (Dörr, 2016). No caso particular do desporto, tema deste trabalho, acresce o facto de ser uma especialidade que desperta o interesse de público de todas as idades e que se presta à utilização de *templates* (van Dalen, 2012), facilitando a automatização.

Em 2010, o The Big Ten Network lançou um serviço deste género recorrendo a um algoritmo da Narrative Science (Lohr, 2011) que produzia automaticamente notícias de basebol usando dados estatísticos. Em 2014, a Associated Press (AP) passou a fazer a cobertura da Minor League Baseball (MLB), acompanhando os resultados de 142 equipas distribuídas por 13 ligas estaduais. Em vez das centenas de jornalistas que seriam necessários para esta missão, a AP recorreu à Automated Insights, para automatizar a produção de notícias com base nos dados estatísticos que já detinha. No mesmo ano, na Europa, o grupo sueco de media locais Östgöta Media automatizou a escrita dos resultados dos jogos de futebol de âmbito local, com o objetivo de aumentar a cobertura noticiosa (Diakopoulos, 2019). Em 2016, a cobertura dos jogos olímpicos no Rio de Janeiro começou a ser relatada por algoritmos, com o *The Washington Post a* utilizar o seu Heliograf na produção automática de conteúdos. Em 2018, o jornal espanhol online nativo El Confidencial lançou o seu algoritmo Anafut com o objetivo de produzir notícias sobre jogos de divisões secundárias de futebol. A partir de 2021, o grupo de media local holandês NDC começou a usar algoritmos na produção de notícias desportivas relacionadas com o futebol amador, propondo-se fazer a cobertura de 60 mil jogos numa temporada. Para atingir este objetivo, a NDC recorreu à empresa tecnológica sueca United Robots, que desenvolveu um algoritmo para combinar os dados estruturados fornecidos pelas associações de futebol com informações enviadas por treinadores ou dirigentes (crowdsourcing) em resposta a uma mensagem enviada pelo sistema. A NDC acredita que a produção automática de textos sobre futebol amador permitirá cobrir jogos sem atenção mediática, atraindo novas audiências que a médio-prazo se poderão tornar assinantes.

Apesar deste trabalho ser sobre jornalismo desportivo, vale a pena recordar que a produção automática de texto também ocorreu noutros campos, nomeadamente na economia e noutras especialidades onde existem muitos dados públicos.

No primeiro caso, logo em 2011 a Forbes recorreu à empresa Narrative Science para automatizar a produção: com o apoio do algoritmo Quill, passando a oferecer relatórios automáticos sobre os resultados das empresas, usando informação obtida nos relatórios de contas disponibilizados na Internet (Dörr, 2016). A Associated Press (AP) também começou a desenvolver experiências neste campo e, em 2014, passou a disponibilizar relatórios sobre o desempenho das empresas usando o algoritmo Wordsmith, da Automated Insight (Graefe, 2016). Também na Europa começaram a surgir sistemas automáticos para a produção de notícias financeiras: neste ano de 2014, a empresa alemã TextOn desenvolveu um algoritmo para produzir notícias sobre economia destinadas a vários jornais do seu país (Dörr, 2016).

E a produção automática não se limitou a estes temas. Em 2011, o Los Angeles Times lançou um espaço online denominado Homicide Report para publicar notícias semiautomáticas e mapas usando dados recolhidos nas esquadras locais de polícia (Young e Hermida, 2015). Três anos depois, o mesmo jornal desenvolveu um algoritmo (Quakebot) para produzir informação sobre uma nova temática de grande interesse na região: tremores de terra.

As experiências iniciais de automatização começaram a surgir para que os media pudessem responder a um público mais exigente em termos de volume de notícias. Os constrangimentos económicos que impossibilitavam novas contratações acabavam por limitar os jornalistas à cobertura de uma parcela das histórias em agenda, sendo obrigados a escolher apenas determinados tópicos (Tandoc *et al.*, 2022). Por isso, a Inteligência Artificial Generativa veio responder parcialmente a esta limitação, mas acabou por criar igualmente algumas preocupações relacionadas com a eventual substituição dos humanos por algoritmos. Porém, o objetivo do uso da IA nas redações não é substituir jornalistas, mas sim ajudar nas tarefas repetitivas que não geram receitas (Cardoso e Baldi, 2021), como a transcrição de entrevistas ou a redação de textos rotineiros.

4. Metodologia

4.1. Perguntas de investigação e técnicas

A revisão bibliográfica conduziu a três perguntas de investigação relacionadas com a produção automática de textos, que neste caso vão ser respondidas no contexto da análise a um algoritmo desenvolvido em Portugal:

PI_1: Os algoritmos usam uma linguagem repetitiva?
PI_2: Os textos produzidos pelos algoritmos precisam de verificação?
PI_3: Os algoritmos poderão roubar trabalho aos jornalistas?

Para responder às duas primeiras perguntas, optou-se pela análise de conteúdo, uma técnica que permite fazer a interpretação dos dados recolhidos de forma sistemática para descrever ou explicar um determinado fenómeno. Usando-se a ferramenta Voyant Tools foram comparadas as características dos textos automáticos com as das notícias publicadas nos media. A amostra dos textos automáticos foi ainda analisada manualmente para verificar se cumpriam as técnicas de redação mais elementares no jornalismo.

Por fim, para responder à terceira pergunta de investigação, foi contado o número de sínteses de jogos produzidas pelo Prosebot e o número de notícias escritas por jornalistas no mesmo intervalo de tempo, procurando-se perceber se existe concorrência ou complementaridade.

4.2 Amostra e ferramentas

Para a análise de conteúdo foi recolhida uma amostra de dez sínteses escritas pelo algoritmo Prosebot, do jornal Zerozero.pt, e de dez notícias sobre o mesmo jogo publicadas em diversos media e plataformas, nomeadamente o Sporting Notícias, o Diário de Santo Tirso, a Fundação Portuguesa de Futebol, o Jornal do Fundão, O Mirante e o Jornal de Mafra. A escolha destes jornais resulta de uma pesquisa em que se procuraram notícias relacionadas com jogos sobre os quais o Prosebot também produziu conteúdos. O reduzido número de notícias que integra a amostra está relacionado com a dificuldade em identificar jogos que no período em análise (1 de maio a 31 de outubro de 2022) tenham sido objeto de síntese do Prosebot e, simultaneamente, notícia nos media.

A ferramenta escolhida para a análise foi a Voyant Tools, por ser totalmente gratuita e por usar várias bibliotecas de código aberto em várias línguas, incluindo o Português. O Voyant Tools é um software utilizado internacionalmente e está constantemente a ser atualizado, o que lhe confere uma maior precisão e eficácia na análise de textos. Comparado com outros softwares de análise de texto, como o Sobek ou o Word Tree, o Voyant Tools tem a mais-valia de permitir calcular parâmetros como a densidade vocabular e o índice de legibilidade, que interessam particularmente nesta investigação.

Com recurso à ferramenta escolhida, analisou-se o corpo de texto de cada síntese e de cada notícia correspondente, excluindo os títulos e os antetítulos. A ferramenta calculou a densidade vocabular (quociente entre número de palavras únicas e o número total de palavras), a média de palavras por frase, e o índice de legibilidade (dificuldade de leitura de um texto obtida através da sua complexidade).

Num segundo momento, os textos passaram por uma avaliação das normas jornalísticas, para identificar as regras de construção de notícias e detetar erros de ordem gramatical ou ortográfica. A análise ficou constituída por um bloco quantitativo e um bloco qualitativo (Tabela 1), a partir dos quais foi feita uma comparação entre os conteúdos gerados pelo algoritmo e as notícias escritas por jornalistas.

4.3. Estudo de caso: Prosebot/Zerozero

O Prosebot é um software de processamento de informação sobre resultados de jogos de futebol. Este sistema foi desenvolvido pela empresa ZOS que gere uma base de dados sobre jogos, jogadores e equipas, com cerca de 5 milhões de entidades (Fernandes, 2021). A base de dados começou a ser construída em 2003, através da participação de colaboradores voluntários que não são jornalistas. Estes colaboradores estabelecem um vínculo

de confiança com o jornal, que lhes dá acesso à base de dados e lhes permite adicionar informações sobre cada jogo. Através desta colaboração, a base de dados do Zerozero.pt "está em contínua ascensão, criando novas oportunidades para apresentação da informação de forma a facilitar a sua interpretação" (Cardoso, 2022, p.ii).

Tabela 1. Categorias na análise de conteúdo das notícias

Bloco quantitativo (Voyant)	Densidade vocabular
	Média de palavras
	Índice de legibilidade
Bloco qualitativo	Texto (estilo, links, citações, ortografia e gramática)
	Níveis de informação

Fonte. Elaboração própria.

O Prosebot é uma ferramenta baseada em Geração de Linguagem Natural (GLN), que funciona com modelos de frases pré-elaborados, um conjunto de condições que ativam esses modelos e ajustes de ordem gramatical, que permitem a concordância entre o género e o número (Fernandes, 2021). A construção dos modelos de texto foi feita em colaboração com os jornalistas do zerozero.pt, através da participação num inquérito.

O algoritmo estrutura o texto em sete partes, descritas por Fernandes (2021): 1. Título: resultado da partida; 2. Antetítulo: caracterização resumida do resultado em função da diferença de golos, número e contexto de golos marcados e equipas participantes; 3. Lead: informações sobre o resultado da partida, os jogadores que marcaram golo, o número de cartões vermelhos e os jogadores relevantes na partida; 4. Introdução: informações sobre o resultado da partida, se a equipa vem de vitórias ou derrotas, o melhor jogador da partida e uma curiosidade sobre o jogo. 5. Eventos: momentos mais relevantes durante o jogo, com destaque para os golos, cartões vermelhos, substituições e penáltis. 6. Próximas partidas: classificação de cada equipa após a partida, o próximo jogo de cada uma e originalidades relacionadas com as equipas; 7. Curiosidades: esta última secção surge separada do texto e inclui uma lista de curiosidades sobre o jogo ou sobre as equipas: geralmente refere-se a séries de vitórias, séries de golos marcados, estatísticas e recordes.

Atualmente, o Prosebot pode ser utilizado em diversas línguas e está também a ser utilizado para antevisões de jogos. Mais recentemente, o código foi disponibilizado em acesso aberto, o que deu "origem a um sistema particularmente útil a criadores de conteúdos e empresas que pretendem acelerar o processo de redação de notícias e sínteses, recorrendo à geração automática de versões semifinalizadas de texto a partir de dados estruturados" (Cardoso, 2022, p.ii).

5. Resultados

Os dados obtidos (Tabela 2) mostram que existem diferenças entre os textos escritos pelo algoritmo e os textos escritos por jornalistas. Os que são produzidos pelo Prosebot apresentam menor densidade vocabular (média de 0,63 vs 0,67), sendo que a densidade é tanto maior quanto mais próxima estiver do valor 1. Esta constatação confirma um dado obtido por Clerwall (2014) num estudo similar em que os participantes consideraram o texto automático como mais descritivo, mas também mais aborrecido devido à repetição vocabular. Neste parâmetro, há quatro notícias do Prosebot que apresentam maior densidade do que as produzidas pelos media, o que aparenta um relativo equilíbrio. O facto de os modelos de frases do Prosebot terem sido elaborados pelos jornalistas do Zerozero.pt poderá justificar esta proximidade dos valores.

De modo geral, os textos produzidos pelo Prosebot apresentam orações mais curtas (média de 17 vs 24,8 palavras), havendo apenas uma exceção em que o algoritmo produziu orações mais longas do que o jornalista (notícia 10).

Tabela 2. Dados obtidos através da ferramenta Voyant

	Densidade vocabular		**Média de palavras**		**Legibilidade**	
	Prosebot	Media	Prosebot	Media	Prosebot	Media
1	0.74	0.722	18.7	27	10.245	10.469
2	0.676	0.772	13.6	30.8	8.313	11.747
3	0.618	0.55	16.4	25.8	9.879	12.63
4	0.623	0.868	17.1	17.7	9.86	10.875
5	0.531	0.7	14.2	20	8,677	9.03
6	0.634	0.661	17.9	27.6	8.517	11.142
7	0.589	0.661	17.9	29.6	8.879	11.566
8	0.63	0.647	15.4	26.1	8.168	9.698
9	0.625	0.599	16	27	9.556	9.844
10	0.672	0.562	23.2	16.4	8.581	9.38

Fonte. Elaboração própria.

No caso do índice de legibilidade, os textos automáticos mostraram-se menos legíveis (média de 9,19 vs 10,63), havendo novamente uma exceção (notícia 5).

Sob o ponto de vista qualitativo, verifica-se que a amostra dos textos automáticos cumpre a estrutura de um texto jornalístico, respondendo às questões da unidade base de informação (o quê, quem, quando e onde) segundo o modelo de pirâmide deitada (Canavilhas, 2006). Porém, um olhar mais crítico conclui que a referência temporal (quando) é sempre "neste sábado" ou "neste domingo'', não aparecendo o dia/mês/ano na página da síntese. A data surge apenas na ficha de jogo (hiperligação que se apresenta no início de cada notícia).

O nível de explicação da pirâmide deitada, onde se incluem as duas restantes perguntas (como e o por quê), é um nível omisso neste tipo de notícias. Considerando que o resultado de um jogo raramente exige o "por quê" entende-se esta ausência, mas o mesmo não sucede com o "como" uma vez que é um dado relevante para explicar a forma como foi marcado um golo, por exemplo.

Embora o nível anterior não exista, as notícias automáticas incluem o nível de contextualização da pirâmide deitada, materializado através das hiperligações internas, que acrescentam informações, como: a) ficha de jogo; b) classificação geral do campeonato; c) equipa inicial dos próximos jogos; d) jogadores em destaque na partida. No que diz respeito a erros ortográficos ou gramaticais, os textos automáticos apresentam imprecisões pontuais (notícia 9, repetição do "o AC") e não têm citações, algo que outros algoritmos já oferecem, como acontece no software desenvolvido pela United Robots para o grupo neerlandês NDC.

A análise qualitativa aos conteúdos dos media mostra que têm uma estrutura de notícia, respondendo a algumas das questões da unidade base de informação. No entanto, não têm nível de explicação, nem de contextualização, e também não apresentam qualquer hiperligação. As referências temporais são semelhantes aos textos do algoritmo: "nesta tarde", "hoje", com a exceção da notícia 8, d'*O Mirante*, que refere o dia "10 de setembro". Na maioria dos casos, a data pode ser encontrada no início ou no final da página. Também neste caso, o estudo parece confirmar uma tendência identificada em estudos anteriores (Kieslich *et al*., 2021) em que os inquiridos não percecionaram qualquer aumento da qualidade do texto jornalístico produzido por IA quando comparado com o produto de jornalistas humanos.

Uma particularidade dos conteúdos dos media é que estes são, na maioria das vezes, acompanhados por fotografias e, num dos casos, por um vídeo (notícia 5), algo que não acontece nas sínteses produzidas pelo algoritmo. Deve ainda ser referido que nenhum dos textos automáticos apresentava qualquer erro ortográfico ou gramatical, o que é apresentado como uma vantagem da IA.

6. Conclusões

A produção automática de notícias existe há mais de uma década e meia, mas só nos últimos anos se assistiu a uma utilização mais intensa. Esta situação pode ser vista como uma tentativa de resposta à crise que levou ao despedimento de milhares de jornalistas, procurando-se formas de aumentar a cadência informativa com menos recursos humanos.

Neste trabalho procurou-se analisar o caso de um algoritmo português –o Prosebot– que desde 2021 produz publicamente textos para a publicação desportiva online Zerozero.pt, comparando a sua performance qualitativa com o trabalho desenvolvido por jornalistas.

Uma das vantagens dos humanos sobre as máquinas é a criatividade, por isso a investigação procurava saber se os algoritmos usam uma linguagem diversificada ou se repetem a mesma estrutura. No caso do Prosebot, os resultados indicam que os textos têm orações mais curtas, menor densidade vocabular e menores índices de legibilidade do que os textos escritos por humanos, obedecendo ainda a uma mesma estrutura sintática. Apesar desta constatação, deve ser referido que a Machine Learning permite desenvolver algoritmos com capacidade de aprendizagem baseada na utilização, falando-se até no treino de algoritmos (Linden, 2017) como uma nova atividade para os jornalistas.

A segunda pergunta de investigação procurava saber se os textos produzidos pelo robot precisam de verificação. A análise apenas identificou imprecisões pontuais, não registando erros ou discrepâncias que indiquem a necessidade de uma revisão feita pelos jornalistas. Uma vez que o funcionamento do algoritmo resulta de uma colaboração permanente entre os programadores e os jornalistas, a estrutura dos textos e as regras foram definidas pela redação, pelo que é menos variada, com se viu, mas tem maior eficiência graças ao trabalho conjunto (Junior, 2011).

Por fim, a última questão procurava saber se os algoritmos poderão substituir os jornalistas. Embora as empresas que desenvolvem estes programas informáticos digam que a sua função é libertar os jornalistas das notícias repetitivas para que se possam dedicar a trabalhos mais exigentes e aprofundados, o receio de despedimentos de jornalistas é uma realidade (Wölker e Powell, 2021). Neste sentido, para responder à questão procurámos comparar as quantidades de textos automáticos e de textos humanos num determinado período: no intervalo temporal em que foi recolhida a amostra de notícias (1 de maio a 31 de outubro de 2022), o algoritmo produziu 37 120 sínteses de jogos ocorridos nesse período, enquanto os jornalistas da redação escreveram 12 919 conteúdos de variados géneros jornalísticos e de maior investigação e contextualização. Verifica-se que a produção de sínteses automáticas é cerca de três vezes superior ao número de conteúdos escritos pelos jornalistas da redação, mas a diversidade de temas tratados pelos jornalistas é muito mais ampla e os trabalhos sobre jogos importantes são sempre produzidos por humanos. Neste sentido, é possível interpretar que o algoritmo é uma ajuda muito relevante e pertinente, já que permite produzir mais textos em menos tempo e garantir a cobertura de um número muito maior de jogos, mas nos temas de maior destaque continuam a ser os jornalistas a produzir os conteúdos.

Para além disso, o reduzido número de notícias que faz parte da amostra (10), e que pode ser considerada uma limitação do estudo, está relacionada com a dificuldade em identificar jogos que no período em análise tenham sido objeto de síntese do Prosebot e notícia noutros media que não o Zerozero. Esta particularidade implicou denominar o artigo de "estudo exploratório", mas permitiu igualmente confirmar que o trabalho dos algoritmos incide fundamentalmente sobre jogos de menor importância mediática, não sendo por isso uma ameaça para o trabalho dos jornalistas. Apesar de os números nos permitirem extrair esta conclusão, futuramente serão desenvolvidos novos estudos mais espaçados no tempo, de forma a confirmar esta tendência.

Embora a crescente capacidade dos algoritmos permita um volume de produção que excede em muito as capacidades de produção humana (Carlson, 2015), a criatividade ainda é um fator diferenciador uma vez que os algoritmos tendem a repetir abordagens e estruturas noticiosas, falhando em elementos fundamentais como a explicação dos acontecimentos ou fenómenos. Esta particularidade, poderá mover o jornalismo em direção à contextualização, respondendo ao "como" e ao 'por quê' (Sirén-Heikel *et al*., 2022) como forma de valorizar os conteúdos humanos. O problema poderá ser o crescente interesse dos usuários nas notícias curtas, os chamados "snacks informativos" (Molyneux, 2018), algo que os algoritmos conseguem fazer com maior eficácia, por isso é importante continuar a estudar as tendências de consumo informativo das novas gerações.

Referências

Beckett, C. (2019). New powers, new responsibilities. A global survey of journalism and artificial intelligence. Recuperado a 19 de dezembro de 2022, de https://bit.ly/42pwo26

Cardoso, N. (2022). Development of an Open-Source Data-to-Text System. [Dissertação de Mestrado, Universidade do Porto]. Recuperado em 19 de dezembro de 2022, em https://bit.ly/3OXIPiH

Cardoso, G. e Baldi, V. (2021). *Algoritmos e notícias-A oportunidade da inteligência artificial no jornalismo*. Relatório OberCom- Observatório da Comunicação. Recuperado a 8 de janeiro de 2023, de https://bit.ly/3a20F1t

Carlson, M. (2015). The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority. *Digital Journalism, 3*(3), 416–431. https://doi.org/10.1080/21670811.2014.976412

Canavilhas, J. (2006). Web journalism: from the inverted pyramid to the tumbled pyramid. Recuperado a 22 de janeiro de 2023, de https://www.bocc.ubi.pt/pag/canavilhas-joao-inverted-pyramid.pdf

Canavilhas, J., Satuf, I., Luna, D., e Torres, V. (2014). Jornalistas e tecnoatores: dois mundos, duas culturas, um objetivo. *Revista Esferas,* 5, 85-95. Recuperado a 7 de janeiro de 2023, em https://bit.ly/38OOJjI

Carreira, K. e Squirra, S. (2017). Jornalismo automatizado, geração de linguagem natural e a lógica do bom suficiente. *Revista Observatório*, *3*(3), 60-84. Recuperado a 21 de janeiro de 2023, em https://10.20873/uft.2447-4266.2017v3n3p60

Chadwick, A., e Vaccari, C. (2019). News sharing on UK social media: misinformation, disinformation & correction. Loughborough University, Online Civic Culture Centre. Recuperado a 10 de dezembro de 2022, de https://hdl.handle.net/2134/37720

Clerwall, C. (2014). Enter the Robot Journalist. Users' perceptions of automated content. *Journalism Practice, 8*(5), 519-531. https://doi.org/10.1080/17512786.2014.883116

De Mateo, R., Bergés, L. e Garnatxe, A. (2010). Crisis, what crisis? The media: business and journalism in times of crisis. *tripleC, 8*(2), 251-274. https://doi.org/10.31269/triplec.v8i2.212

Diakopoulos, N. (2019). *Automating the news: how algorithms are rewriting the media.* Harvard University Press.

Dörr, K. N. (2016). Mapping the field of Algorithmic Journalism. *Digital Journalism, 4*(6), 700-722. https://doi.org/10.1080/21670811.2015.1096748

Edwards, C., Autumn E., Patrick R. S., e Ashleigh K. S. (2014). Is that a Bot Running the Social Media Feed? Testing the Differences in Perceptions of Communication Quality for a Human Agent and a Bot Agent on Twitter. *Computers in Human Behavior*, 33, 372–376. https://10.1016/j.chb.2013.08.013

Fernandes, P. (2021). Community-based Sports Articles Generation Platform using NLG and Post-Editing. [Dissertação de Mestrado, Universidade do Porto]. Recuperado a 7 de janeiro de 2023, de https://hdl.handle.net/10216/135617

Graefe, A. (2016). *Guide to Automated Journalism. Tow Center for Digital Journalism*. https://bit.ly/45VLdN4

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism, 7*(8), 993-1012. https://doi.org/10.1080/21670811.2019.1623700

Junior, W. (2011). Jornalismo computacional em função da "Era do Big Data". *Líbero, 14*(28), 45-52.

Kieslich, K., Došenović, P., Starke, C., Lünich, M., e Marcinkowski, F. (2021). Artificial Intelligence in Journalism. How does the public perceive the impact of artificial intelligence on the future of journalism? *Factsheet, 4.*

Kim, D. e Kim, S. (2018). Newspaper journalists' attitudes towards robot journalism. *Telematics and Informatics, 35*(2), 340-357. https://doi.org/10.1016/j.tele.2017.12.009

Linden, C. (2017). Decades of automation in the newsroom still so many jobs in journalism? Why are there still so many jobs in journalism. *Digital Journalism*, *5*(2), 123-140. https://doi.org/10.1080/21670811.2016.1160791

Lohr, S. (2011). *In Case You Wondered, a Real Human Wrote This Column. The News York Times.* Recuperado a 5 de dezembro de 2022, de https://nyti.ms/3tG1RgK

Molyneux, L. (2018). Mobile News Consumption. *Digital Journalism*, *6*(5), 634-650.

Murcia Verdú, F. J., Ramos Antón, R. e Calvo Rubio, L. M. (2022). Análisis comparado de la calidad de crónicas deportivas elaboradas por inteligencia artificial y periodistas. *Revista Latina de Comunicación Social,* 80, 91-111. https://doi.org/10.4185/RLCS-2022-1553

Napoli, P. (2014). On Automation in Media Industries: Integrating Algorithmic Media Production into Media Industries Scholarship. *Media Industries Journal, 1*(1), 33-38. https://doi.org/10.3998/mij.15031809.0001.107

Pena, P. (2019). *Fábrica de mentiras. Viagem ao mundo das fake news*. Penquim Random.

Peña-Fernández, S., Meso-Ayerdi, K., Larrondo-Ureta, A. e Díaz-Noci, J. (2023). Without journalists, there is no journalism: the social dimension of generative artificial intelligence in the media. *Profesional de la Información*, 32(3), e320227. https://doi.org/10.3145/epi.2023.mar.27

Pérez-Seijo, S., e Vicente, P. N. (2022). After the hype: how hi-tech is reshaping journalism. Em Jorge Vázquez-Herrero, Alba Silva-Rodríguez, María-Cruz Negreira-Rey, Carlos Toural-Bran e Xosé López-García (Ed.): *Total Journalism: Models, Techniques and Challenges* (pp. 41-52). Springer International Publishing.

Quandt, N., Sant'Anna, R., Winques, K. e Máximo, M. (2021). Análise de apurações jornalísticas feitas com uso de Inteligência Artificial. *Redes-Revista Interdisciplinar do IELUSC,* (4), *39-52*. Recuperado a 18 de janeiro de 2023, de https://bit.ly/3GnGdo7

Schwab, K. (2016). *The Fourth Industrial Revolution*. Random house.

Sirén-Heikel, S., Leppänen, L., Lindén, C. G., e Bäck, A. (2019). Unboxing news automation. *Nordic Journal of Media Studies, 1*(1), 47-66. https://doi.org/10.2478/njms-2019-0004

Tandoc Jr., E. C., Wu, S., Tan, J., e Contreras-Yap, S. (2022). What is (automated) news? A content analysis of algorithm-written news articles. *Revista Media & Jornalismo*, 22(41), 103-120. https://doi.org/10.14195/2183-5462_41_6

Torrijos, J. (2021). Semi-automated Journalism. *News Media Innovation Reconsidered: Ethics and Values in a Creative Reconstruction of Journalism*, 124-137.

Túñez-López, J. M., Fieiras Ceide, C. e Vaz-Álvarez, M. (2021). Impact of Artificial Intelligence on Journalism: transformations in the company, products, contents and professional profile. *Communication & Society, 34*(1), 177-193. https://doi.org/10.15581/003.34.1.177-193

Van Dalen, A. (2012). The algorithms behind the headlines. How machine-written news redefines the core skills of human journalists. *Journalism Practice, 6*(5-6), 648-658. https://doi.org/10.1080/17512786.2012.667268

Ventura-Pociño, P. (2022). *Algorithms in the newsrooms: Challenges and recommendations for artificial intelligence with the ethical values of journalism*. Catalan Press Council (CIC). Recuperado a 8 de janeiro de 2023, de https://t.ly/6XL

Wölker, A. e Powell, T. (2021). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, 22, 86-103. https://doi.org.10.1177/1464884918757072

Young, M. L. e Hermida, A. (2015) From Mr. and Mrs. Outlier to central tendencies, *Digital Journalism*, *3*(3), 381-397. https://dx.doi.org/10.1080/21670811.2014.976409

Breve CV dos autores

João Canavilhas é doutor em "Comunicação, Cultura e Educação" pela Universidade de Salamanca, DEA em Comunicação Audiovisual e Publicidade pela mesma instituição e Licenciado em Comunicação Social pela Universidade da Beira Interior. Atualmente é professor na Universidade da Beira Interior e investigador no Labcom – Comunicação e Artes. A sua investigação centra-se em vários aspetos da relação entre o jornalismo e as novas tecnologias.

Adriana Gonçalves é doutoranda em Ciências da Comunicação na Universidade da Beira Interior, Mestre em Jornalismo e Licenciada em Ciências da Comunicação pela mesma instituição. É bolseira de doutoramento no Labcom- Comunicação e Artes, onde investiga o potencial das tecnologias de produção de texto no campo do jornalismo.

Declaração de autoria CrediT

Conceitualização: J.C., A.G.; Metodologia: J.C., A.G.; Análise formal: J.C., A.G.; Redação (revisão e edição): J.C.; A.G.; Visualização:A.G.; Supervisão: J.C.

Conflito de intereses

Os autores declaram não haver qualquer conflito de interesses.