LITERATURA CIENTÍFICA SOBRE A MINERAÇÃO DE TEXTOS APLICADA À IDENTIFICAÇÃO DA PERSONALIDADE DE ATLETAS

Resumo - Um dos desafios das pesquisas científicas é de proporcionar objetividade a partir dos dados obtidos também por meio de depoimentos e entrevistas. Na pesquisa Memórias Olímpicas por Atletas Olímpicos Brasileiros, do Grupo de Estudos Olímpicos da Universidade de São Paulo, foram realizadas aproximadamente 1400 entrevistas com atletas olímpicos brasileiros, resultando em um acervo de informações fundamental para o entendimento do esporte olímpico brasileiro. Na etapa do projeto descrito neste manuscrito, objetivou-se buscar, por meio de uma pesquisa bibliográfica não exaustiva, diferentes estudos que realizaram extração de dados, a partir de métodos associados à mineração de dados, por meio da pesquisa na base do Google Acadêmico, utilizando as palavras-chave: ("text mining" ou "mineração de dados") + (atleta ou esporte) + (psicologia ou personalidade). Foram analisados 67 resultados. Ainda que os resultados indiquem publicações que apresentam os descritores da busca, observa-se que a maioria apenas fez uso dos termos de maneira teórica ou introdutória, mas sem aplicar efetivamente algum formato de extração de textos. Nesse sentido, destaca-se a importância de conhecer o que as produções acadêmicas têm apresentado a respeito da mineração de textos sobre personalidade e aspectos emocionais, presentes em pesquisas associadas a atletas ou no contexto esportivo.

Palavras-chave: Mineração de texto; text mining; atleta; personalidade; esporte.

SCIENTIFIC LITERATURE ON TEXT MINING APPLIED TO THE IDENTIFICATION OF THE ATHLETES´ PERSONALITY

Abstract – One of the challenges of scientific research is to provide objectivity from the data obtained also through testimonials and interviews. In the Olympic Memories by Brazilian Olympic Athletes project applied by the Olympic Studies Group of the University of São Paulo, 1400 interviews were conducted with Brazilian Olympic athletes, resulting in a collection of information fundamental to the understanding of the Brazilian Olympic sport. In the project stage described in this manuscript, the objective was to study, through a non-exhaustive initial bibliographical method, search different studies that applied methods of data extraction, from methods associated with data mining, through the Google base search Academic, using the keywords: ("text mining" or "data mining") + (athlete or sport) + (psychology or personality). 67 results were analyzed. Although the results indicate publications that present the descriptors proposed in the search, it is observed that the majority only used the terms in a theoretical or introductory way, or even superficially in relation to the object of study but did not effectively apply some format of extraction of texts. In this sense, the importance of knowing what the academic productions have presented regarding the mining of texts about personality and emotional aspects, present in research associated with athletes or in the sport context, is important.

Keywords: Text mining; athlete; psychology; personality; sport.

LITERATURA CIENTÍFICA SOBRE MINERÍA DE TEXTOS APLICADA A LA IDENTIFICACIÓN DE LA PERSONALIDAD DE LOS ATLETAS

Resumen - Uno de los desafíos de la investigación científica es proporcionar objetividad a partir de los datos obtenidos también a través de testimonios y entrevistas. En el proyecto Memorias Olímpicas de Atletas Olímpicos Brasileños aplicado por el Grupo de Estudios Olímpicos de la Universidad de São Paulo, se realizaron 1400 entrevistas con atletas olímpicos brasileños, lo que resultó en una recopilación de información fundamental para la comprensión del deporte olímpico brasileño. O objetivo fue estudiar, a través de un método bibliográfico inicial no exhaustivo, buscar diferentes estudios que aplicaran métodos de extracción de datos, desde métodos asociados con la extracción de datos, a través de la Búsqueda de base académica de Google, utilizando palabras clave: ("minería de texto" o "minería de datos") + (atleta o deporte) + (psicología o personalidad). Se analizaron 67 resultados. Si bien los resultados indican publicaciones que presentan los descriptores propuestos en la búsqueda, se observa que la mayoría solo usó los términos de forma teórica o introductoria, o incluso de manera superficial en relación con el objeto de estudio, pero no aplicó efectivamente algún formato de extracción de textos. En este sentido, la importancia de saber qué han presentado las producciones académicas con respecto a la minería de textos sobre la personalidad y los aspectos emocionales, presentes en las investigaciones asociadas con atletas o en el contexto deportivo, es importante.

Palabras-clave: Minería de textos; deportista; psicología; personalidad; atleta.

Ivan Sant´Ana Rabelo

Escola de Educação Física e Esporte

Universidade de São Paulo

ivanrabelo@usp.br

Katia Rubio

Escola de Educação Física e Esporte

Universidade de São Paulo

katrubio@usp.br

http://dx.doi.org/10.30937/2526-6314.v2n1.id37

Introdução

Narrativas biográficas e histórias de vidas são métodos de coleta de dado qualitativos presentes em diferentes pesquisas acadêmicas. No caso do esporte, conforme destacado por Rubio[1], quando os atletas referem-se a sua trajetória esportiva apresentam-se em suas narrativas a lembrança de pessoas e profissionais que influenciaram e determinaram o desejo pelo esporte, pela busca de melhores condições de vida e de treinamento ou a convivência com outros atletas que também competiam naquele momento histórico e cujas carreiras se cruzaram apontando para a necessidade premente de contextualizar essas situações para promover o entendimento de episódios marcantes de suas vidas e de seus resultados. Em se tratando de atletas de modalidades coletivas essa condição é ainda mais evidente porque vários deles narram suas memórias sobre um mesmo conteúdo vivido a partir de diferentes pontos de vista, apresentando novos conteúdos, uma nova história, apontando para a subjetividade que envolve a construção e elaboração desse tema ainda que vivido coletivamente.

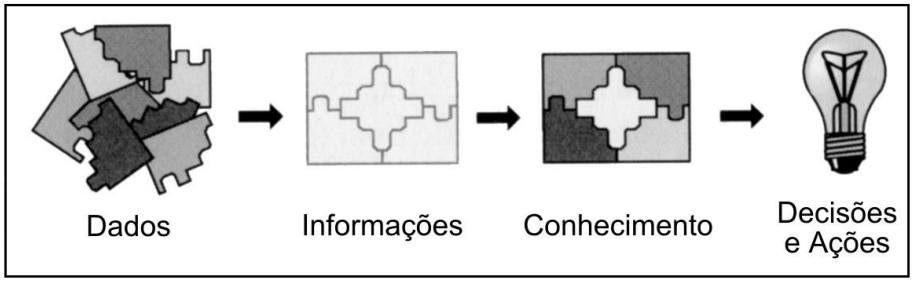
Flusser[2] (p. 97) destaca a distinção entre comunicação dialógica e comunicação discursiva, na qual, para produzir informação, os homens trocam diferentes informações disponíveis na esperança de sintetizar uma nova informação, se referindo à forma de comunicação dialógica. Enquanto que, com o fim de preservar, manter a informação, os homens compartilham informações existentes na esperança de que elas, assim compartilhadas e compartilhadas, possam resistir melhor ao efeito entrópico da natureza, tratando-se da forma de comunicação discursiva. Porém, uma das grandes dificuldades é "produzir diálogos efetivos, isto é, de trocar informações com o objetivo de adquirir novas informações". Sobretudo quando o volume de informações é muito alto, dificultando ainda mais a interpretação destas informações contidas em narrativas discursivas, seja com o fim de produzir e descobrir novas informações, seja com o fim de manter estas memórias ao longo do tempo.

A mineração de dados é parte do processo de *Knowledge Discovery in Databases* (KDD), em português Descoberta do Conhecimento em Banco de Dados. O termo mineração de dados é utilizado como sinônimo para todo o processo de descoberta de conhecimento, contudo, apesar de ser uma etapa muito importante, responde por 15 a 25 por cento do processo de descoberta. De maneira que a mineração

de dados é um processo de extrair informações válidas antes desconhecidas, de grandes bases de dados, auxiliando em decisões cruciais no mundo dos negócios[3].

Nesse sentido, a mineração de dados ou data mining, utiliza-se de técnicas ou algoritmos de áreas, entre elas, aprendizado de máquinas, estatística, redes neurais, algoritmos genéricos etc. Está apoiada no conhecimento indutivo, permitindo descobrir novas regras e padrões nos dados minerados. Kantardzic[4] já apontava que o processo de mineração de dados é composto por cinco fases: definição do problema; seleção, coleta dos dados; pré-processamento dos dados; especificação de possível método; interpretação e análise dos dados produzidos pelo método. Os métodos ou técnicas de mineração de dados são, na verdade, algoritmos computacionais. Cada um desses algoritmos tem características particulares, normalmente entradas e saídas específicas.

Figura 1. Transformando dados em conhecimento para a tomada de decisão



Fonte: Oliveira Neto, 2003

Conforme observado na Figura 1, a coleta dos dados e posterior análise dos dados permite a integração do conhecimento extraído, transformando isso em informação e conhecimento. Antes da tomada de decisão, a mineração vasculha esse conjunto de dados para melhorar seus processos de aprendizagem e descoberta de conhecimento. Diferentes técnicas de data mining são fortemente desenvolvidas para atuar tanto com questões do passado e do presente (análise retrospectiva), e até mesmo em antecipação de comportamentos e eventos, na análise de previsões[5].

Avançando para uma parte mais específica da mineração de dados, voltada a análise de textos, sendo dados textuais escritos em língua natural, ou seja, dados não estruturados, o processo de extração de conhecimento é denominado de mineração de textos ou *text mining*. Desta maneira, a mineração de textos pode ser entendida como

uma especificação da mineração de dados, correspondendo à aplicação de um conjunto de técnicas para descoberta de conhecimento inovador em textos[6].

Autores esclarecem que no processo de mineração de textos pode ser compreendido como uma sequência de etapas genéricas, que devem ser determinadas de acordo com os dados disponíveis e o conhecimento que se espera encontrar nestes dados. Observam-se a existência de diferentes possibilidades de aplicações desse processo são diversas, como, por exemplo, análise exploratória de conteúdo textual, organização de coleções de documentos, análise de sentimentos e sistemas de recomendação, assim como, em análises exploratórias de documentos, que podem possibilitar a identificação e o mapeamento de padrões no conteúdo presentes em dados textuais[7,8,9,10,11].

De acordo com Pietroforte[12] independentemente do contexto da aplicação ou da técnica utilizada, a natureza não estruturada dos dados traz grandes embates para o processo de mineração de textos, pois, nas línguas naturais, as palavras podem exprimir diferentes relações que podem interferir no significado dos textos, entre eles, por exemplo, a sinonímia, a hiperonímia, a polissemia e ambiguidade. As relações semânticas entre palavras e sentenças podem impactar em como as pessoas interpretam os textos e, por isso, podem ser importantes para a mineração de textos.

O entendimento de textos escritos em língua natural é um processo complexo, que se dá por meio do conhecimento das palavras e de seus significados, das relações existentes entre as palavras, bem como do conhecimento de mundo e do contexto no qual o texto foi escrito. E nesse contexto, recursos e técnicas de processamento de língua natural podem auxiliar o tratamento de textos. Entre as análises sintática e semântica de textos que podem ser realizadas automaticamente pode-se citar a identificação de classes morfossintáticas (como substantivos e verbos) e a identificação de papéis semânticos (como os agentes de ações descritas por verbos) para as palavras presentes em documentos[13]. E assim, busca-se empregar técnicas de mineração de textos para apoiar a análise de documentos, possibilitando a identificação de padrões em narrativas e sua relação com traços de personalidade.

De acordo com Marques et al.[14], existem algumas técnicas utilizadas em mineração de textos. Uma delas, denominada Técnica Regra de Associação (*Association*

Rule) permite recuperar todos os padrões interessantes em uma base de dados. E como se a base de dados fosse uma coleção de transações.

Outra técnica é denominada de Agrupamento (*Clustering*). Uma dada população de eventos ou novos itens podem ser particionados (segmentados) em conjuntos de elementos 'padrões'[3]. No mesmo sentido, Bramer[15] corrobora que os algoritmos de agrupamento analisam os dados para encontrar grupos de itens que são semelhantes. Amostras de agrupamento são representadas como um vetor de medições, ou, mais formalmente, como um ponto em um espaço multidimensional. As amostras de um agrupamento válido são mais semelhantes (não necessariamente iguais) entre si do que as amostras que pertencem a um agrupamento diferente.

Na Técnica Dados em Séries Temporais (*TimeSeries Data*) padrões podem ser encontrados em posições de uma série temporal de dados, que é uma sequência de dados capturada a intervalos regulares, por exemplo, segundos, horas, dias, semanas, etc.[3]. Para Han e Kamber séries temporais consistem em sequências de valores ou medidas repetidas excessivamente em intervalos de tempos[16].

Os autores explicam que na análise de séries temporais existem dois objetivos: modelagem das séries temporais, que propõe levantar uma visão sobre os mecanismos subjacentes ou forças que geram as séries temporais; e previsão de séries temporais, que objetiva predizer os futuros valores do tempo de séries de variáveis. As técnicas mais usadas são a análise de tendências e a pesquisa por semelhança[16].

Verifica-se também a técnica de Padrões sequenciais (*Sequential Patterns*), no qual Elmasri e Navathe apontam que a técnica de padrões sequenciais é a investigação de sequências de ações ou eventos[3]. Han e Kamber escrevem que é uma técnica desafiadora, pois pode gerar e/ou testar um número combinatório explosivo de sequências intermediárias[16]. Essa técnica tem algumas similaridades com a técnica de regras, a diferença é que faz exame da dimensão sequencial dos dados analisados, ainda que, por vezes, não seja possível ser identificada essa sequência padrão de acontecimentos. De maneira que, utilizando a técnica de padrões sequências, esse comportamento padrão seria percebido no resultado da mineração.

Conforme apontado por Tan[17], T*ext mining* surgiu como uma área derivada do D*ata mining*, que extrai de textos, a partir de técnicas e processos, informação útil sem precisar de leitura prévia. Por meio do *Text mining* é possível extrair informação

desconhecida de grandes coleções textuais, sem que haja necessidade da leitura humana para extração de dados. De acordo com o autor, podem ser descobertos padrões e relações entre os textos, que poderia ser muito complexo ou difícil para textos com alto volume de informações, podendo muitas vezes serem quase impossíveis de serem analisados com uma leitura manual.

A ideia é de que, para o tratamento das palavras, deverão ser realizadas atividades de pré-processamento comuns em mineração de textos, por meio de uma limpeza dos textos como remoção de pontuações e números, bem como a remoção de *stopwords*. A remoção de *stopwords* visa a eliminação de palavras que não trazem informação relevante para o processo de mineração de textos, de acordo com os objetivos estabelecidos. Essas palavras, chamadas de *stopwords*, normalmente são palavras que possuem as funções de artigos, preposições, pronomes e conjunções. No entanto, também podem ser identificadas *stopwords* específicas do domínio de aplicação do processo, ou seja, palavras que sabidamente são frequentes na coleção e que não distinguem classes ou grupos que se espera identificar com a mineração de textos[13].

Para a identificação automática de verbos e adjetivos realiza-se a anotação morfossintática. A anotação morfossintática é uma tarefa de processamento de língua natural na qual é realizada a atribuição de etiquetas morfossintáticas às palavras de uma sentença. Essas etiquetas identificam a função sintática de cada palavra, tais como substantivo, verbo, adjetivo e preposição. Essa é uma tarefa importante do processamento de língua natural e que serve como base a várias outras tarefas. Após a normalização é realizada a contagem automática da frequência dos termos, sejam eles palavras, verbos ou adjetivos normalizados. Com essa contagem, gera-se uma matriz de frequências para cada abordagem[13] (Rubio et al., 2019).

Sobre o processo de influência do narrador, na construção do seu discurso, segundo Martins et al.[18] a sumarização, em geral, é uma atividade bastante comum. Quando se narra um evento a uma pessoa, costuma-se fazer um resumo do que aconteceu e não uma narração completa e detalhada. Inconscientemente, as pessoas estão sempre sumarizando, quer oral, quer textualmente. Exemplos de sumários escritos incluem notícias de jornais, artigos de revistas, resumo de textos científicos, entre

muitos outros. Por sua utilidade e frequência, há um grande interesse em automatizar esse processo[18].

Ainda a este respeito, de acordo com Martins at al.[18] a sumarização automática vem sendo explorada desde a década de 50, quando começaram a surgir os primeiros métodos para a produção de extratos, sendo o método das palavras-chave[18] (Luhn*, 1958 apud Martins et al., 2001, p. 8) o mais significativo então. Entretanto, como métodos "cegos", fazendo uso de técnicas superficiais, os resultados apresentavam inúmeros problemas, quer de coesão, quer de coerência[18] (Hutchins†, 1987 apud Martins et al., 2001, p. 8), razão pela qual a área ficou praticamente estagnada nas décadas seguintes, voltando a ser objeto de interesse com o advento da *internet* e, portanto, com o aumento considerável de documentos disponíveis on-line e com a necessidade de se "digerir" informações em larga escala, isto é, em grande quantidade e no menor tempo possível.

Como exemplos, no campo da Computação Afetiva, pesquisadores estudam a importância de considerar os estados afetivos e emocionais. Dentre estas características, a personalidade na formação de grupos de aprendizagem, embora, de acordo com Reis et al.[19], poucos são os trabalhos que apresentam os reais impactos de considerar os estados afetivos dos alunos na aprendizagem em grupo.

Nesta pesquisa de Reis et al.[19], por exemplo, foi realizado um mapeamento sistemático, para investigar quais e como os estados afetivos, a personalidade, são considerados na formação de grupos em ambientes de Aprendizagem Colaborativa com Suporte Computacional (CSCL - *Computer Supported Collaborative Learning*). Entre os principais resultados, destaca-se que 16 estudos (76,2%) consideraram os traços de personalidade na formação de grupos em ambientes CSCL, e grande parte desses estudos estão relacionados à detecção de estados afetivos dos aprendizes em ambientes CSCL via preenchimento de questionário pelo respondente. Além disso, embora 8 estudos (38,1%) incluam uma investigação empírica, os resultados obtidos pela comunidade científica sobre afetividade na formação de grupos em ambientes CSCL ainda são incipientes, havendo muitas oportunidades para novas pesquisas.

* Luhn HP. The automatic creation of literature abstracts. IBM Journal of Research and Development. 1958; 2: 159-165.

† Hutchins J. Summarization: Some problems and Methods. In: Meaning J. The frontier of informatics. Cambridge: London; 1987. p. 151-173.

Segundo Peres[20] uma abordagem léxica, assim como, uma hipótese léxica, estão relacionadas à uma perspectiva ainda em desenvolvimento na psicologia da personalidade, presentes em modelos teóricos com necessidade de mais estudos aplicados, porém, trata-se de uma perspectiva a partir da qual alguns dos principais modelos teóricos da investigação da personalidade foram desenvolvidos, como o modelo de Cattell dos 16 fatores primários e o modelo dos cinco fatores ou *Big Five*. Tais abordagens fundamentam-se na ideia de que a maioria das características da personalidade socialmente relevantes e salientes estão codificadas na linguagem natural das diferentes culturas ao longo de sua história, em termos descritores de traços da personalidade podem ser retirados dos léxicos dos idiomas.

No mesmo sentido de pesquisas de personalidade à partir de modelos teóricos, pesquisas analisando a personalidade de indivíduos são encontradas na literatura, em diversos domínios, sobretudo no contexto da Avaliação Psicológica, tais como as pesquisas apresentadas em[13,21,22,23,24,25,26,27,28], entre outros, muitas destas analisando traços de personalidade de amostras do contexto esportivo. Ao mesmo tempo, os poucos estudos de análise da personalidade com atletas estão mais fortemente vinculados ao uso de medidas padronizadas em formato de lápis-papel, ou testagem com uso de testes psicológicos informatizados, *biofeedback*, porém, ausência de estudos específicos de análise de aspectos emocionais e de personalidade, à partir de entrevistas com atletas.

Além disso, mesmo que pareça ser muito forte a presença da tecnologia, de inovação nas mais distintas ciências, especificamente no contexto esportivo, verificam-se estudos que analisam dados de maneira automática, porém, muito focadas nos aspectos biológicos, fisiológicos, de desempenho, mas ainda raros do ponto de vista de aspectos emocionais, que tenham extraído as variáveis de maneira automática, em atletas e equipes. Corroborando, a seguir será apresentado um estudo com análises feitas com algoritmos em aprendizagem de máquina.

Portanto, nesse exemplo supracitado, foi aplicado a busca de informações de atletas, por meio de uma técnica denominada de *SCOUT* Voleibol em que, conforme Andrade[29], objetiva-se capturar e processar informações estatísticas de desempenho dos atletas da equipe e da equipe adversária, monitorando assim desempenhos de atletas nos fundamentos (saque, defesa, ataque, bloqueio, recepção). Permitindo, com isso, auxiliar na tomada de decisão, na elaboração de planos técnicos (saque, levantamento, cortada

etc.) e táticos (como evitar os tipos de jogadas do adversário). Entretanto, a grande quantidade de dados gerados, muitas vezes desnecessários, dificultam a visualização de padrões de jogo. Isso também força que a análise seja realizada após as partidas[29].

A ferramenta *Waikato Enviroment Knowledge Analysis* (WEKA) é destacada por Marques et al.[14], que explica se tratar de uma coleção de algoritmos da aprendizagem de máquina para tarefas de mineração de dados. Tais algoritmos podem ser aplicados diretamente a um conjunto de dados. O WEKA contém ferramentas para o pré-processamento dos dados, a classificação, a regressão, o agrupamento, as regras da associação e visualização, se mostrando também adequado para o desenvolvimento de novos sistemas de aprendizagem[30].

Contudo, conforme será apontado nos resultados da busca que será apresentada neste manuscrito, são raros os estudos que atuaram especificamente utilizando-se de análises de discursos à partir de entrevistas, com métodos de extração automática dos dados, neste caso em específico, fazendo uso de métodos associados à mineração de textos. Restringindo ainda mais se nos referirmos à aspectos da personalidade, emoções, e outros aspectos psicológicos de atletas e esportistas, como o proposto neste artigo.

Por fim, do ponto de vista das contribuições sobre o tema, por meio de revisões de literatura, antes mesmo de iniciar a revisão, os pesquisadores realizam buscas nas bases de dados acadêmicos, a fim de melhor definir a questão de pesquisa, avaliar a viabilidade da revisão e obter maior familiaridade com o tema. A etapa de busca para revisão sistemática, entretanto, deve ser mais criteriosa, seguindo-se procedimentos padronizados e com o devido registro do que se fez, pois isso possibilitará a sua reprodução. Como o objetivo da revisão sistemática é apresentar a síntese da evidência disponível sobre uma questão de pesquisa, a busca e a seleção devem ser bem executadas, de modo que seja possível identificar e incluir estudos relevantes sobre o assunto[31].

Objetivos

O objetivo deste artigo foi buscar, por meio de uma pesquisa bibliográfica não exaustiva, diferentes estudos que aplicaram métodos de extração de textos, de aspectos da personalidade em narrativas de atletas. Como objetivo específico, buscou-se (1) levantamento bibliográfico para a compreensão dos assuntos envolvidos na pesquisa sobre Processamento de Linguagem Natural, Mineração de Texto e Análise de aspectos

psicológicos; (2) realização de bibliografia quantitativa com o objetivo de levantar o número de trabalhos existentes relacionados ao assunto; (3) estudo e apresentação de sínteses de alguns dos trabalhos com dados relacionados ao tema deste manuscrito, juntamente com um quadro comparativo destacando algumas características de classificação dos resultados encontrados.

Portanto, buscam-se respostas para as seguintes perguntas de pesquisa: (a) existem publicações associadas a estudos de extração de texto, de maneira automática em amostras de atletas ou esportistas, sobre características psicológicas e de personalidade? (b) qual a quantidade de sujeitos contidos nos grupos amostrais de tais estudos? Para tal, realizou-se uma revisão da literatura, seguindo as recomendações apresentadas em Kitchenham[32] e envolvendo as seguintes etapas: definição da busca, execução da busca e extração e análise dos resultados.

A hipótese da pesquisa é de que seja possível identificar características associadas às emoções, sobretudo, de traços de personalidade, a partir da aplicação de técnicas de mineração de textos nas entrevistas realizadas com atletas, porém, há pouca literatura científica a este respeito.

Método

Foi realizada uma revisão da literatura acerca dos estudos disponíveis que fizeram uso de mineração de dados para levantamento de características afetivas e emocionais no contexto esportivo. Trata-se de uma pesquisa qualitativa e exploratória quanto a abordagem teórico-metodológica.

Foi consultada a base de dados Google Acadêmico (Scholar Google), sendo utilizados os seguintes termos de busca com operadores booleanos: ("text mining" OR "mineração de texto") AND (personalidade OR psicologia) AND (esporte OR atleta). A escolha pela base Google Acadêmico se deu justamente por ser considerada uma base "muito abrangente", segundo Pereira e Galvão[31], Ferenhof e Fernandes[33] entre outros, haja vista se tratar de uma busca inicial, com intuito exploratório a respeito do tema, podendo estimular pesquisas futuras em outras bases de periódicos específicos.

A busca na base do Google Acadêmico não estipulou limites de data. O primeiro critério de inclusão para os estudos que retornaram foi apresentar alguma utilização da mineração de texto ou do termo "*text mining*", associado aos demais booleanos. Para

realizar a pesquisa, inicialmente foram analisados os títulos e resumos dos estudos, não sendo identificado estudos duplicados na base dos resultados retornados.

Posteriormente, foi proposto uma nova seleção por critérios de exclusão. Nessa etapa, foram excluídos resultados de Patentes e Citações (com este critério, diminuiu de 83 para 67 resultados). Em seguida, foram identificados os estudos dos quais não foi possível recuperar o texto completo.

Para cada resultado selecionado, portanto, foram analisados se tratou-se de uma pesquisa com coletas diretas em seres-humanos ou utilizando dados de fontes, tais como metadados, postagem em redes sociais etc. Além disso, buscou-se também analisar o ano das publicações. Todos os procedimentos supracitados de busca, foram realizados durante o mês de janeiro de 2019.

Entende-se que este manuscrito contribui como uma das etapas iniciais e determinantes para o projeto de pesquisa de análise de aspectos psicológicos, de personalidade, por meio da mineração automática de textos, associada ao projeto Memórias Olímpicas por Atletas Olímpicos Brasileiros.

Resultados e Discussão

Como ponto de partida para esta pesquisa, foi estabelecida a utilização do mecanismo de busca Google Acadêmico para o levantamento de artigos relacionados aos termos aplicados. Assim, a busca na referida base de dados retornou um total de 83 resultados associados aos termos propostos no mecanismo de busca.

De acordo com Nunes[34], em virtude de seu algoritmo, o Google Acadêmico apresenta resultados de suas buscas por ordem de relevância. Na palestra citada, Nunes explica que as cinco primeiras páginas de resultados devem, portanto, cobrir os achados merecedores de maior foco de atenção, de maneira satisfatória.

Apesar deste entendimento parecer adequado, em razão da quantidade mais rebaixada de publicações recuperadas na busca com estes booleanos, foram analisadas todas as 7 páginas de resultados, totalizando os resultados retornados nesta busca. De forma que o aspecto da "relevância" não foi abordado, já que todo o universo de respostas da busca foi considerado.

Ressalta-se que os termos utilizados na busca, tendo sido os operadores booleanos: ("text mining" OR "mineração de texto") AND (personalidade OR psicologia) AND (esporte OR atleta), foram escolhidos, sobretudo, acreditando

representarem palavras importantes para a referida pesquisa, já que, entre os exemplos de combinações possíveis com os booleanos acima, incluirão Psicologia do Esporte, Personalidade de atleta, Avaliação psicológica em atletas, entre outras combinações, sempre agregado também ao termo "mineração de texto", ou em inglês "text mining", que são os termos específicos desta pesquisa.

Abrindo a caixa preta, os pesquisadores, também tiveram como possibilidade inicial, incluir o termo "aprendizagem de máquina" e seu correspondente em inglês "*machine learning*", porém, em uma varredura inicial, verificou-se que em buscas com estes dois termos incluídos, retornavam muitos resultados mais fortemente associados a extração de dados, e não especificamente por meio de textos, que é o método que se pretende estudar em maior profundidade, já que está relacionado diretamente ao projeto com entrevistas de atletas, capturadas em formato de texto nesta fase da pesquisa.

Assim, no critério a) patentes e citações retornou a exclusão de 16 resultados, rebaixando de 83 para 67 resultados; quanto ao critério b) estudos dos quais não foi possível recuperar o texto completo, foi removido inicialmente 1 resultado por não ser possível o acesso ao manuscrito, por se tratar de um resultado de natureza Diretório de Ciência Desportiva, enumerando a construção de ementas, e não uma pesquisa ou produção de conhecimento científico, e ainda, o manuscrito não apresentava informações suficientes para responder aos objetivos desta revisão de literatura; também verificou-se 2 resultados em formato de Livro, que também não permitiram o acesso direto. Na tabela 1 é possível verificar dados quantitativos a partir da busca realizada.

Tabela 1. Dados quantitativos a partir dos resultados apresentados

Tipo	Qtde	%
Dissertação	27	40
Tese	12	18
Artigo	10	15
Anais	4	6
Monografia	3	4,5
TCC	3	4,5
Livro	2	3
Relatório técnico	2	3
Simpósio	2	3
Diretório	1	1,5
Relatório pós-doutorado	1	1,5
Total Geral	67	

Verifica-se a predominância de resultados no formato de Dissertação de Mestrado, com 40% dos resultados, seguida de 18% de resultados associados à Teses de Doutorado. Totalizando portanto, mais de 50% da produção de estudos com os referidos marcadores booleanos, verificam-se que os estudos ainda se mostram mais associados à produção dentro de programas de pós-graduação, ainda que, espera-se que a produção destas teses e dissertações possam produzir à posteriori um aumento no número de artigos, haja vista que entre os objetivos dos programas de pós-graduação está na disseminação do conhecimento produzido pelos orientandos dentro das universidades e levar este conhecimento à disposição da comunidade, por meio da publicação de artigos, assim como, apresentações de trabalhos em eventos científicos que, de acordo com a tabela 1, reforçam a presença de 9%, se somado os Anais e Simpósio, resultados da busca. Os artigos, que são uma fonte importante de divulgação de produção científica, retornou 10 resultados (15%) nesta revisão bibliográfica.

Ainda em relação a tabela 1, nota-se também um número somado de 6 resultados (9%) associados a produção nos programas de Graduação, Cursos de Extensão ou Pós-graduação lato-sensu que são os formatos de Monografias e Trabalhos de Conclusão de Cursos (TCC). Em um nível de pós-doutoramento, nesta busca, retornou apenas um único resultado (1,5%).

Quanto ao acesso, dos 67 resultados recuperados, um número de 61 deles tinham o acesso gratuito ao conteúdo completo; 2 resultados possibilitavam o acesso pago, sendo um artigo dos EUA/Canadá em inglês e outro que se tratava de uma Dissertação de Mestrado em italiano. Além disso, recuperou-se 2 resultados que se tratavam de livros, portanto, não permitindo o acesso completo ao conteúdo.

Ainda assim, optou-se por revelar quais os tipos de estudos predominantes nos dados revelados pela busca. Na tabela 2 é possível analisar a presença dos grupos amostrais com base nos resultados retornados.

Tabela 2. Contextos das pesquisas retornadas pela busca

Palavras-chave	Contextos	Frequência	%
Atleta ou esporte	Contexto esportivo ou amostras com atletas	11	16
	Não específicos	56	84
Personalidade ou psicologia	Investigações sobre personalidade ou aspectos psicológicos	18	27
	Não específicos	49	73
"*Text mining*" ou "mineração de texto"	Específicos sobre mineração de textos	46	69
	Não específicos	21	31

Observou-se que a palavra "atleta" ou "esporte" estava presente como critério de entrada para a busca dos resultados, porém, quando analisados os conteúdos destes resultados retornados, o que se observa é que a maioria dos resultados apresentavam estes dois ou um dos dois booleanos de maneira apenas presente na introdução, ou em citações de informações referindo-se a outras pesquisas, mas sem haver qualquer menção mais explicita ou efetivamente direcionada no método informações de dados amostrais com atletas.

No que tange aos resultados apontados em relação aos termos booleanos "personalidade" e/ou "psicologia", também são observados resultados que utilizam tais temas de maneira muito superficial, ou mesmo, em sentido diferente ao proposto no artigo, por exemplo, referindo-se à "personalidade da marca" (referindo-se as características de uma empresa de marketing), ou mesmo "personalidade da mídia", referindo-se à uma pessoa famosa. Portanto, nesse sentido, estes marcadores booleanos, assim como atleta e/ou esporte, ainda que tenham sido colocados como critérios de busca, em sua maioria, não estão diretamente relacionados ao objeto desse estudo.

Também foram verificados dentre os resultados da busca, estudos que investigaram extração de dados de pesquisas com não-humanos (animais, plantas etc.), em um número de apenas 1 artigo dentre os resultados da busca, com amostras de serpentes. Na tabela 3 é possível analisar a presença dos grupos amostrais com base nos resultados retornados.

Tabela 3. Grupos amostrais das pesquisas retornadas pela busca

Grupos amostrais	Frequência	%
Humanos	17	25,5
Dados indiretos	49	73
Animais	1	1,5

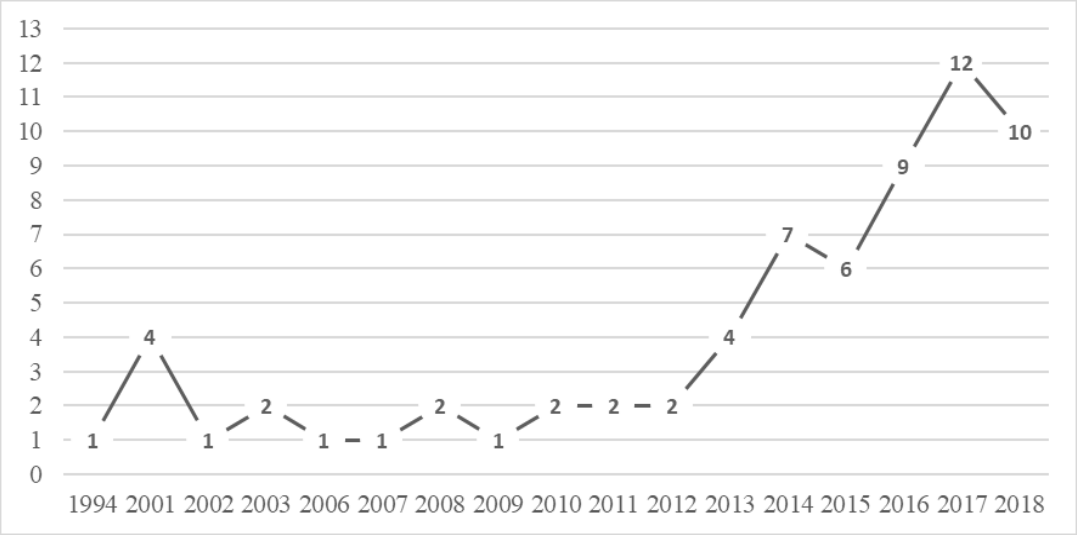
Quanto ao ano de publicação dos resultados reportados, observou-se publicações realizadas de 1994 a 2018. Estes dados podem ser verificados na tabela 4 e também no gráfico da figura 2.

Tabela 4. Ano das pesquisas retornadas pela busca

Ano	Frequência	%
1994	1	1,5
2001	4	6
2002	1	1,5
2003	2	3
2006	1	1,5
2007	1	1,5
2008	2	3
2009	1	1,5
2010	2	3
2011	2	3
2012	2	3
2013	4	6
2014	7	10,5
2015	6	9
2016	9	13
2017	12	18
2018	10	15
Total	67	

O gráfico que descreve os anos de publicações dentre os resultados reportados, de 1994 a 2018, permite observar um aumento na quantidade de publicações à partir do ano de 2013.

Figura 2. Gráfico dos anos das pesquisas retornadas pela busca



No sentido do histórico de estudos sobre o tema mineração de textos de maneira automática, de acordo com o que é apresentado na pesquisa de Silva[35], recuperada por esta revisão de literatura, desde o final dos anos 80, um grande esforço de pesquisa vem sendo desenvolvido com o intuito de se extrair padrões úteis e desconhecidos a partir do grande volume de dados existentes nas organizações. A primeira vertente de pesquisa explorou, principalmente, dados estruturados. Mais recentemente, passou-se a dar mais atenção a dados na forma de texto. Entretanto, passados alguns anos de pesquisa em Mineração de texto, observa-se que esse tipo de tecnologia ainda é pouco explorado. Considerando que a maior parte das informações disponíveis está em forma textual e que nessa forma podem estar escondidos padrões importantes questiona se o porquê da pouca utilização de mineração de texto.

Esse resultado que se refere a Silva[35], ainda que se mostre repleto de conteúdo muito importante para o entendimento a respeito de estudos e métodos de Mineração de textos, contudo, não é específico do campo do esporte, ou com amostragens que incluam atletas. Ao mesmo tempo, assim como algumas das citações presentes neste manuscrito que não são específicas do contexto do esporte, ou mesmo de áreas associadas a psicologia, ainda assim serão brevemente destacadas em alguns trechos, acreditando contribuir para o conhecimento que poderá ser ponderado e, sendo analisado com cautela, transposto para o contexto esportivo, ou mesmo, base ou instigarem pesquisas parecidas, replicadas com atletas e esportistas.

Assim, relatando alguns dos demais artigos retornados pela busca, que atendiam todos os critérios booleanos propostos, no artigo de Vissoci et al.[36] o estudo analisou a influência do esporte na formação da identidade de 25 atletas de Futsal (25,49 ± 4,91 anos e 9,12 ± 3,59 anos de prática) das equipes participantes da Liga Nacional de Futsal, por meio da caracterização do padrão semântico do discurso e do relato de história de vida. Por meio dos resultados destacou-se a presença de três grupos caracterizados pela expressão de suas metamorfoses identitárias. O primeiro grupo apresentou uma forma de racionalidade instrumental, ligados à vitória, sucesso e outras características do esporte espetáculo. Os dois grupos mais autonômicos uniram experiências esportivas e não esportivas, alcançando racionalidade comunicativa e multiplicidade de papéis em manifestações de personagens. Os autores concluíram que o esporte pode possibilitar fragmentos emancipatórios dependendo da multiplicidade e riqueza de papéis e estimulações ambientais, favorecendo pluralidade de experiências que instiguem transições ecológicas de autonomia.

O método de análise foi do tipo dedutiva, partindo de categorias pré-definidas de classificação dos atletas que foram categorizados de acordo com o conteúdo direcionados para autonomia ou heteronomia e com os projetos de vida, fundamentados em políticas de identidade ou em identidades políticas[36] (Ciampa‡, 1987 apud Vissoci et al., 2018, p. 59). Os grupos foram: (a) discurso voltado para heteronomia e que não evidenciam um projeto de vida; (b) discurso voltado para autonomia com um projeto de vida fundamentado em políticas de identidade; (c) discurso voltado para autonomia e com um projeto de vida fundamentado em identidades políticas. Cada grupo foi caracterizado dentro do processo de metamorfose-emancipação da formação identitária através de história de vida e da trajetória esportiva ilustradas com relatos de sujeitos emblemáticos e padrão semântico das palavras com análise em rede. Análise em Redes. Técnicas de mineração de texto foram aplicadas para criação de um corpus com base nas entrevistas. As entrevistas eram mineradas, por parágrafos, na seguinte sequência: a) remoção de conectores entre as palavras; b) redução de verbos para radicais (sistematização); c) remoção de pontuações; d) revisão manual do corpus para limpeza. Foi calculada a ocorrência de cada palavra em cada parágrafo e uma matriz de associação utilizando correlação de Spearman.

‡ Ciampa AC. A estória do Severino e a história da Severina. São Paulo: Brasiliense; 1987.

Cada palavra do corpus caracterizou um nodo na rede conectado pelas hastes, que representam a intensidade da sua correlação. Nessa rede bipartite, palavras que fossem expressas nos mesmos parágrafos teriam uma relação maior (direta), com relações indiretas estabelecidas pela presença de palavras em comum. Por exemplo, ao dizer "Eu gosto de futsal", os termos "gosto" e "futsal" teriam uma associação alta e direta. Contudo, se em outro parágrafo ao relatar "Eu gosto de vencer", "futsal" e "vencer" teriam uma associação indireta pelo compartilhamento da sentença com a palavra comum "gosto". Quanto mais próximos os nodos, maior é associação entre as palavras. O descritivo de rede *betweness* (conectores) foi utilizado para identificar as palavras que eram mais importantes para fazer conexões entre os clusters da rede de conexões semânticas. Todas as análises foram conduzidas com o programa Linguagem R[36] (RProject[§], 2014 apud Vissoci et al, 2018, p. 60).

Sobre a Compreensão da formação identitária, a análise do processo de identidade-metamorfose-emancipação foi feita por meio do método de história de vida a partir do relato da história da carreira atlética. Para cada grupo de atletas foi selecionado um sujeito emblemático que evidenciasse momentos determinantes de possiblidades emancipatórias ou colonizadoras oferecidas pelo contexto durante as transições, sentido atribuído pelos atletas e movimentos de metamorfose identitária[36] (Ciampa, 1987 apud Vissoci et al, 2018, p. 59). A análise foi descrita através de relatos de história de vida, utilizando nomes fictícios aos sujeitos emblemáticos.

Aproveitando da oportunidade criada pelos Jogos Olímpicos de 2016, sediado no Brasil, a pesquisa de Ramos[37], que tratou de uma grande captura de dados textuais em português e inglês da rede social Twitter, foi investigado os motivos que levam a comunidade falante do português e a comunidade falante do inglês a manifestar desejos em relação a pessoas associadas aos Jogos Olímpicos. Utilizando métodos de processamento de texto em linguagem natural, mineração de textos para encontrar os desejos, análise de sentimento para classificação de desejos e técnicas de refinamento para exposição dos desejos foi possível levantar fatores que podem motivar desejos.

Nesta pesquisa de Ramos[37] a fonte de dados para o trabalho foi a hashtag oficial dos jogos no Twitter: #Rio2016, considerando os idiomas Português e Inglês. Foi

[§] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna: Austria; 2009 [acesso 5 mai. 2019]. Disponível em: http://www.R-project.org/.

possível observar que desejos positivos são fruto de contínuo cumprimento de expectativas geradas pelos fãs para determinado atleta e que apenas uma expectativa não cumprida pode levar a desejos negativos para um atleta por dias. Fatores de desejo não relacionado a atletas também foram observados como atenção da mídia, política e nacionalidade. Principais verbos relacionados aos desejos realizados durantes os Jogos Olímpicos também foram levantados. Caso haja uma complexa junção de diferentes sentimentos na mesma frase o algoritmo tende a classificar como neutro do que assumir que a frase é polarizada. Dessa forma boa parte dos desejos é classificado como neutro, mas considerando apenas os pólos positivo e negativo é possível perceber que existem mais desejos positivos do que desejos negativos.

Apesar de um desejo poder ser tão complexo quanto a mente humana foi possível desenvolver um processo que aplica técnicas computacionais avançadas de processamento de linguagem natural e de mineração de desejos para capturar indícios de desejos relacionados aos Jogos Olímpicos que pode abrir uma série de aplicações interessantes na área de Sistemas de Informação[37].

Um artigo que despertou curiosidade, ainda que não se tenha o acesso ao artigo completo e os dados de quantidade de participantes e outras informações necessárias para alcançar o objetivo desta revisão de literatura, mas que mostrou-se relevante, parcialmente atendendo a alguns dos booleanos, mas não específico do contexto do esporte ou com atletas, trata-se do estudo de Fengler, e Frozza[38], conforme publicado nos Anais do Salão de Ensino e de Extensão, do trabalho "Classificação de sentimentos e sua validação em comentários de usuários", que traz em seu resumo que a Internet mostra-se para além de um ambiente de pesquisas, mas sobretudo é evidente se tratar de um local para as pessoas expressarem suas opiniões sobre diversos assuntos, como uma notícia ou comentário sobre um produto adquirido.

Segundo os autores, as opiniões que antes eram divididas com amigos e familiares, hoje também são compartilhadas na web podendo influenciar, na compra de um produto em uma loja virtual. Portanto, torna-se determinante acompanhar os comentários dos usuários, o que gera uma grande quantidade de dados, que aumenta a cada dia, impedindo a realização de uma análise não automatizada desses dados. A área de Análise de Sentimentos busca identificar o sentimento que o texto expressa e classificar a polaridade do comentário em positivo, negativo ou neutro.

A partir da seleção de comentários de produtos eletrônicos em lojas virtuais, compondo uma base de comentários, os autores objetivaram aplicar diferentes métodos para avaliar a classificação da polaridade de comentários, abordando Análise de Sentimentos, sendo aplicados os métodos de validação: Grupo de Usuários, utilização da Ferramenta IFeel e uso de Redes Neurais Artificiais. O grupo de usuários responderam um questionário para opinem se os comentários eram positivos, negativos ou neutros e porque chegaram a cada classificação. Após esses resultados foram comparados aos resultados obtidos pela Ferramenta IFeel e pela Rede Neural Artificial (RNA), utilizando-se a mesma base de comentários. A Ferramenta IFeel é uma ferramenta web, que utiliza diferentes métodos de Análise de Sentimentos. RNA é uma técnica advinda da área da Inteligência Artificial que apresenta um modelo computacional baseado na aprendizagem e no funcionamento do cérebro humano, atuando principalmente no reconhecimento de padrões[38].

No trabalho apresentado em um Simpósio Argentino de GRANdes DAtos, por Coelho, Lima e Omar[39], são destacadas que as investigações em análise automáticas de documentos vêm permitindo grandes avanços permitindo no reconhecimento de aspectos subjetivos. Nesta pesquisa, utilizaram a classificação da polaridade do texto, ou seja, o quão negativa ou positiva são as opiniões expressadas em um texto. Assim, o trabalho teve como objetivo o estudo e a aplicação de uma ferramenta que, conterá um algoritmo de classificação de sentimentos, sendo ele capaz de avaliar a polaridade de comentário extraído de uma base de dados de livros da Amazon e outra do Twitter, baseando-se em técnicas de mineração de textos. A base avaliada coletou 1700 mensagens e foi usado o OpLexicon como base de conhecimento (palavras anotadas como positivo, negative ou neutron) e Sentilex.

Este trabalho apresentou um framework de Classificação de Sentimentos por polaridade (positivo, negativo) para Mineração de Textos capaz de classificar automaticamente essa polaridade em comentários de textos extraídos do Twitter e de um dataset da Amazon. O trabalho também mostrou o diferencial do framework destacando suas principais características. O modelo de classificação de sentimentos do framework permitiu conhecer a opinião da rede social do usuário do Twitter e de usuário da Amazon sobre o livro e o filme "50 Tons de Cinza". As postagens dos usuários foram categorizadas neste trabalho em um sentimento: positivo e negativo. Além disso, nos

experimentos realizados, a determinação do sentimento tanto para relação a uma entidade (lista de entidade pré-definidas), quanto para documento geral é permitido pelo framework sem fazer uma associação a uma entidade específica[39].

Trabalhos futuros envolvem aplicar a ferramenta no contexto de outros idiomas para validar a capacidade de generalização com a troca da base de conhecimento. O próximo passo é trabalhar com textos em português[39] (Souza; Vieira[**], 2012 apud Coelho et al., 2017, p. 1). Assim, os autores corroboram para o entendimento de que, por ser uma área ainda em desenvolvimento, métodos criados para estas análises, na maioria, são para língua inglesa, o que dificulta sua utilização em textos escritos em português.

Vale destacar ainda que neste trabalho[39] é descrito que uma das áreas da mineração de texto também conhecida como mineração de opiniões e análise de sentimentos, busca classificar textos não por tópicos, e sim pelo sentimento ou opinião em documentos. Esses documentos em geral estão associados à classificação binária (0 ou 1) entre sentimentos (positivo e negativos), usa-se o termo de forma abrangente para demonstrar como são tratadas opiniões, sentimentos e subjetividade em documentos textos de forma computacional[40].

Já a análise de sentimentos busca por meio de documentos textuais, opiniões que se formam, mas sem se preocupar no que está sendo comentado, e sim com a opinião constituída no que refere-se a sua polaridade, ou seja, se uma menção por exemplo, é positiva ou negativa com relação a um determinado produto[39,41]. Aplica-se mineração de opiniões em documentos textos em qualquer tamanho e formato, como páginas web, post, comentários, tweets, etc. A opinião é composta por dois elementos chaves, um alvo e um sentimento sobre esse alvo. Um alvo classifica-se em uma entidade, aspecto de uma entidade ou tópico que pode ser representado por um produto, pessoa, organização, marca, evento, etc. Por outro lado, um sentimento representa uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo[39,42].

Guelpeli[43] apresenta em sua tese que o volume de informações não estruturadas vem crescendo de forma desordenada nos dias atuais, e a Internet, representa um repositório destas informações em grandes partes textuais é o grande agente facilitador

[**] Souza M, Vieira R. Sentiment analysis on twitter data for portuguese language. Computational Processing of the Portuguese Language. PROPOR 2012. Lecture Notes in Computer Science, vol 7243. Berlin: Heidelberg; 2012.

de novos conhecimentos. Esta estrutura, quase anárquica, trouxe consigo um grande problema de organização, surgido ante a dificuldade do ser humano de armazenar grande quantidade de informações e depois recuperá-las, ocasionando uma sobrecarga de informação. A área de Descoberta de Conhecimento em Textos, cuja principal finalidade é obter algum tipo de conhecimento em documentos textuais, torna explícito o conhecimento implícito.

Na tese de doutorado de Peres[20] intitulada "The personality lexicon in Brazilian Portuguese: studies with natural language", foi gerada uma lista de descritores da personalidade para o português brasileiro utilizando a rede social Twitter como fonte. Como resultado, acumulou-se 1.454 adjetivos, seis nomes, 10 pronomes e 383 substantivos, potenciais descritores para a construção de uma taxonomia brasileira da personalidade. Em um outro estudo, ainda na referida tese, está relacionado à análise da dimensionalidade de um corpus também obtido no Twitter, com 172 adjetivos e 86.899 sujeitos, cujo resultados sugeriram dois promissores modelos a serem utilizados em futuras pesquisas, um com sete e outro com 14 dimensões. Na tese também são discutidas questões metodológicas e teóricas, destes estudos de linguagem natural para a pesquisa futura em personalidade. De maneira que este estudo se mostra muito relevante para o estudo do qual este manuscrito está se propondo a integrar, que é uma pesquisa para extração de dados a partir da mineração de textos. Contudo, vale informar que o estudo de Peres[20] não é realizado com população de atleta ou mesmo com foco no contexto esportivo.

Ainda nesse mesmo sentido, porém, para além dos artigo capturados por esta revisão de literatura, em outros dois artigos, quais sejam, "Identificação dos Traços de Personalidade de Alunos com Base em Postagens no Facebook" e "Utilização da ferramenta Five Labs para Identificação de Traços de Personalidade dos Estudantes", foi observado que tais publicações se aproximaram um pouco mais das propostas de aplicação de mineração de textos, mais especificamente, em ambos os artigos objetivou-se a extração de traços de personalidade pelo Facebook, por meio do instrumento FIVE LABS. Porém, observou-se ser a mesma coleta de dados, com 49 sujeitos, para ambos os artigos. Mais, ainda que o método seja relativo a extração de dados a partir de textos, reforça-se a ausência de relação do grupo amostral com população de atletas.

Ainda do ponto de vista de uso de ferramentas computadorizadas, em uma outra Tese de Doutorado denominada "Desenvolvimento de uma metodologia de ensino, baseada na teoria das inteligências múltiplas viabilizada pelo uso de tecnologias de informação e comunicação", descreve-se o desenvolvimento de uma metodologia de aprendizagem baseada na teoria das inteligências múltiplas, aplicada em um laboratório virtual em ambiente 3D utilizando o software OpenSim, integrado ao MOODLE. Na dissertação "Tecnologias de informação e comunicação para analisar a responsividade emocional em organizações hierárquicas", objetivou-se explorar o uso de TICs no contexto da comunicação em organizações hierárquicas, provendo métodos de expressão e compreensão emocional a fim de entender comportamentos emocionais.

Também foi encontrado um estudo fora desta revisão de literatura que tratou, de maneira mais próxima, do contexto esportivo, tal como a dissertação "Personal trainer & Cia: noções de marketing na literatura sobre treinamento personalizado", ainda que não tenha analisado amostras específicas de atletas, e o foco não seja analisar características de esportistas, e sim, analisar aspectos associados ao marketing e treinamento esportivo.

Em uma dissertação externa aos dados destes booleanos propostos, mas intitulada "O direito à privacidade e a proteção aos dados pessoais na sociedade da informação: uma abordagem acerca de um novo direito fundamental", objetivou-se verificar na literatura brasileira e estrangeira como o direito à privacidade vem sendo abordado no contexto da proteção aos dados pessoais dos indivíduos, frente aos impactos de uma sociedade da informação e do consumo, de forma que o foco de estudo da publicação relacionou-se à análise de base de dados textuais, mas não em relatos ou coletas com pessoas diretamente.

Já no trabalho de conclusão de curso da graduação em Administração, Mantovani[44], trata da aquisição de informações contidas em redes socais, para a identificação de interesses que possam influenciar usuários a preferirem trajetos alternativos durante o deslocamento urbano. Com a finalidade de atingir o objetivo, foi proposta a coleta de informações da rede social Facebook e uma classificação de interesses a partir dos dados adquiridos para a identificação de interesses. Em seguida, com o intuito de avaliar a classificação realizada, foi selecionado os interesses que possam ter influência na recomendação de trajetos para embasar um questionário

personalizado, a ser respondido pelos indivíduos que forneceram seus dados para a pesquisa. Como resultado, os interesses obtidos a partir de redes sociais de propósito geral podem influenciar a preferência de pessoas em um trajeto baseado em interesses. Especificamente, os interesses relacionados à "Lazer" e "Compras" contribuem para a preferência de trajetos.

Outros resultados externos também se mostravam vinculados a extração e análise de dados a partir de corpus linguísticos, de bases bibliográficas ou acesso a informação via rede sociais, como por exemplo, Facebook e Twitter, mas não atreladas ao contexto esportivo ou com atletas. Em outra publicação, objetivou-se compreender como se deu a formação da identidade dos descendentes dos africanos que vieram para o Brasil como escravos, a partir da análise do conteúdo dos livros didáticos de história do Brasil. Assim como, verificou-se uma pesquisa relacionada a extração de textos, refere-se à uma análise do uso das mídias sociais na atividade de inteligência policial, relacionado ao atentado na maratona de Boston.

Vale ressaltar nesse manuscrito os poucos resultados de publicações internacionais, considerando a utilização do termo em inglês "*text mining*", contudo, ao mesmo tempo, junto a este booleano em inglês, havia a necessidade das palavras atleta ou esporte, assim como as palavras personalidade ou psicologia, de maneira que em razão destes segundo e terceiro critério estar em português, já esperava-se encontrar pesquisas mais frequentes na língua portuguesa, ainda que, tenha sido verificado neste manuscrito, alguns resultados publicados em Simpósio, Anais e outros meios, fora do Brasil, ou em outras línguas, sendo em inglês (6), espanhol (2) e italiano (1), enquanto em português foram 55 artigos e mais 3 artigos em português de Portugal. Ainda que alguns artigos estejam em idiomas estrangeiros, isso deve-se em razão de que em alguma parte do texto, tal como no resumo, incluem estes marcadores booleanos citados, mais associados a língua portuguesa.

Corroborando com estes resultados de ausência de pesquisas a respeito de corpus ou métodos criados para a língua portuguesa, do Brasil, durante as pesquisas para obter o embasamento teórico do trabalho de Fengler e Frozza[38], a respeito da parte sobre a bibliometria, foi constatado que existem poucas ferramentas com métodos para realizar a análise de frases ou textos escritos em língua portuguesa. A maioria das ferramentas é destinada para realizar análise a partir de textos escritos na língua inglesa. Os autores

destacam na pesquisa a necessidade de se realizar a análise de sentimentos em textos construídos na língua portuguesa.

Como limitações de pesquisas de mineração de texto em áreas específicas, com populações específicas e considerando linguagem e cultura, os autores Vissoci et al.[36] também destacam que apesar da pesquisa ter sido realizado com atletas que representassem a população de jogadores de futsal de rendimento, os resultados se restringem às experiências apresentadas podendo ser compreendidos numa perspectiva cultural da modalidade do futsal brasileiro. Contudo, entendemos que a estrutura esportiva em outros países e modalidades podem ser semelhantes e ter resultados similares. Portanto, muitas vezes toma-se a decisão de apresentar poucos casos de sujeitos emblemáticos, seguindo as sugestões da literatura, que possibilitam o uso de casos que retratam claramente o processo teórico estudado.

Outra limitação está na análise de mineração de texto pela falta de dicionários mais adequados para o procedimento à língua portuguesa requerendo algumas limpezas manuais dos corpus. Entretanto, os resultados são corroborados pelos estudos de caso, de forma que sustentam sua argumentação. Como implicações práticas, fica evidente a importância de elementos do contexto, como as transições ecológicas instigadas pelo ambiente, que estimulam modificações no processo de desenvolvimento do indivíduo[36].

Portanto, sugere-se que sejam realizadas revisões de literatura também com abrangência de outros temas e em outras línguas, objetivando verificar a presença de estudos de extração de dados por meio de textos, em amostras de atletas e esportistas. Ao mesmo tempo, reforça-se a importância de que sejam realizados estudos, como este manuscrito, especificamente focadas em populações de atletas, assim como, no contexto esportivo em geral, que considerem o aspecto cultural presente em cada país, que podem estar expressas por meio da linguagem local, coloquial, e portanto, mais sensíveis a extrapolação de resultados de uma cultura para outra, assim como de um país para o outro.

Ainda que este manuscrito tenha se concentrado em uma abordagem automática de extração de informação e conhecimento, enfatiza-se a necessidade da presença humana para guiar o funcionamento de quaisquer destas ferramentas. Considera-se que o papel destas seja de auxiliar na verificação de milhares de registros de dados, seja por comunalidades ou discrepâncias que possam passar desapercebidas a análise de um

pesquisador, ou ainda agilizar o tempo de análise quantitativa e consequentemente qualitativa das informações. Assim, o acompanhamento do analista na interpretação do resultado produzido pela máquina torna-se imprescindível Rubio, Rabelo e Cruz[45], tendo em vista que os softwares não apresentam competências humanas importantes, tais como a intuição, a experiência humana, a imaginação e criatividade, que podem diferenciar aspectos relevantes de correlações irrelevantes, dar sentido ao material minerado.

Tais aspectos ressaltados visam, neste momento, apenas fortalecer a necessidade de estudos de inteligência artificial, aprendizagem de máquina, mineração de dados, mineração de textos em com a população brasileira, para que possam ser verificadas extrapolações que poderiam ser aplicadas, sem adaptações mínimas do conhecimento da cultura e valores de uma população, contexto etc.

Considerações Finais

Ainda que os resultados indiquem estudos que apresentam os descritores propostos na busca, observa-se que a maioria apenas fez uso dos termos booleanos propostos na busca, de maneira teórica ou introdutória, mas não tenha aplicado efetivamente algum método de extração de dados no estudo, sobretudo com atletas ou amostras no contexto esportivo. Nesse sentido, destaca-se a importância de se conhecer o que as produções acadêmicas têm apresentado a respeito da mineração de dados nas ciências humanas e no contexto do esporte, presentes em pesquisas nos mais variados contextos, associados a tecnologia, psicologia, educação física e outras áreas. O que reflete a interdisciplinaridade das pesquisas dentro dos marcadores de busca propostos neste trabalho.

Em conclusão, corroborando com outros autores de épocas distintas, trata-se de um conhecimento que ainda está começando a ser apresentado na literatura acadêmica. Isso é muito motivador e impulsiona a necessidade de mais pesquisas de características emocionais de atletas, utilizando recursos de mineração de dados, já que se observa que a análise de texto tem se mostrado relevante, conforme já tratado por Tan[17] e também por Nunes[46] para compreensão de aspectos afetivos e de personalidade em análises textuais de maneira automática.

Em suma, é fundamental que a ciência se abra para receber as importantes contribuições que emergem das vidas vividas, para o conhecimento que pulsa nas

narrativas orais e que esse conhecimento possa ser devidamente organizado, preservado, revisto e compartilhado com a comunidade em geral. Peirce pode nos dar uma pista valiosa quando afirma que nosso conhecimento nunca é absoluto, mas é como se sempre flutuasse em um *continuum* de incerteza e indeterminação[47].

Por fim, destaca-se que, o desenvolvimento de métodos que ajudem a extrair e organizar os dados, permitirá diferentes tipos de análises, contribuindo para o estudo de associações entre o conteúdo mais frequentemente presente nas narrativas e, a partir de análises fatoriais, o cruzamento de dados com luz às teorias de personalidade e outros cruzamentos de dados. Constituído como um dos principais fenômenos sociais contemporâneos, o esporte tem se estabelecido como um campo privilegiado de estudo e intervenção, seja nos aspectos específicos de sua prática tática e técnica, mas também educativo e sociocultural.

Analisando o estado da arte da área de mineração automática de textos, verificamos que, apesar dos avanços, não existe ainda uma única maneira ou ferramenta. Como observado pelo caminho trilhado neste trabalho, a mineração de textos pode ser tratada em distintas perspectivas e apoiada em variadas técnicas, modelos, seja em nível superficial ou profundo, mas cada qual apresenta méritos e dificuldades. Apoiados nas últimas tendências este trabalho procura contribuir ao desenvolvimento de um método de investigação automática de informação, em entrevistas de atletas olímpicos, com base em metodologias híbridas, voltado para o apoio de pesquisadores, contribuindo no aprimorando dos métodos de avaliação e investigação dos estudos produzidos nesta área.

Referências

1 Rubio K. Memórias e narrativas biográficas de atletas olímpicos brasileiros. São Paulo: Képos; 2014. p. 45-90.
2 Flusser V. O mundo codificado. São Paulo: Cosac Naif; 2007.
3 Elmasri R, Navathe SB. Sistemas de banco de dados. São Paulo: Pearson Addison Wesley; 2005.
4 Kantardzic M. Data mining: Conceps, models, methods, and algotithms. Hoboken, Nova Jersey, EUA: John Wiley & Sons Inc; 2003.
5 Oliveira Neto JSD. Um modelo conceitual de dados voltado para aplicações de CRM baseado em reutilização de atributos. [dissertação]. Santa Catarina: Universidade Federal de Santa Catarina; 2003.
6 Sinoara RA, Camacho-Collados J, Rossi RG, Navigli R, Rezende SO. (2019). Knowledge-enhanced document embeddings for text classification. Knowledge-Based Systems. 2019; 163:955-971. doi:10.1016/j.knosys.2018.10.026.

7 Marcacini RM, Rossi RG, Matsuno IP, Rezende SO. Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. Decision Support Systems. 2018; 114: 70-80. doi:10.1016/j.dss.2018.08.009.
8 Sinoara RA, Antunes J, Rezende SO. Text mining and semantics: a systematic mapping study. Journal of the Brazilian Computer Society. 2017; 23: 9. doi:10.1186/s13173-017-0058-7.
9 Sinoara RA, Scheicher RB, Rezende SO. Evaluation of latent dirichlet allocation for document organization in different levels of semantic complexity. Proceedings of 2017 IEEE Symposium Series on Computational Intelligence (SSCI); 2017. p. 1-8. doi:10.1109/SSCI.2017.8280939.
10 Santos FF. Extração de tópicos baseado em agrupamento de regras de associação. [tese]. São Carlos: Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação; 2015. doi:10.11606/T.55.2015.tde-02122015-161054.
11 Rezende SO. Sistemas inteligentes: fundamentos e aplicações. São Paulo: Editora Manole; 2003.
12 Pietroforte AVS. Semântica lexical. In: Introdução à linguística II: Princípios de análise. São Paulo: Editora Contexto; 2010.
13 Rubio K, Rabelo IS, Sinoara RA, Rezende SO. (2019). Quem procura acha: mineração de textos na identificação da personalidade de atletas olímpicos. Gerais: Revista Interinstitucional de Psicologia. 2019. No prelo.
14 Marques ÉB, Zamberlam ADO, de Oliveira RF, Raimann LH, de Oliveira LV. Projeto de módulo de Data Mining para Scout Voleibol. Seminário de Informática - RS (SEMINFO RS 2008), Torres (RS); 2008.
15 Bramer, M. Principles of Data Mining (Undergraduate Topics in Computer Science). Londres: SpringerVerlag London Ltd; 2007.
16 Han J, Kamber M. Data cining concepts and Techniques. São Francisco: Morgan Kaufman Publishers; 2006.
17 Tan AH. Text mining: The state of the art and the challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, vol. 8; 1999. p. 65-70. Disponível em: <http://textmining.krdll.org.sg>. Acessado em: 19 mai. 2018.
18 Martins CB. Pardo TAS, Espina AP, Rino LHM. Introdução à sumarização automática. Relatório Técnico RT-DC. 2001; 2(1): 35.
19 Reis RCD, Rodriguez CL, Lyra KT, Jaques PA, Bittencourt II, Isotani, S. Estado da arte sobre afetividade na formação de grupos em ambientes colaborativos de aprendizagem. Revista Brasileira de Informática na Educação. 2015; 23(3): 113–130.
20 Peres AJDS. The personality lexicon in Brazilian Portuguese: studies with natural language. [tese]. Universidade de Brasília; 2018.
21 Capretz LF. Ahmed F. Why do we need personality diversity in software engineering? ACM SIGSOFT Software Engineering Notes. 2010; 35(2): 1–11.
22 Bartholomeu D, Machado AA, Spigato F, Bartholomeu LL, Cozza HFP, Montiel JM. Traços de personalidade, ansiedade e depressão em jogadores de futebol. Rev. bras. psicol. Esporte. 2010; 3(1): 98-114.
23 Cruz S, Da Silva F, Monteiro C, Santos P, Rossilei I. Personality in software engineering: preliminar findings from a systematic literature review. Proceedings of 15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011); 2011. p. 1–10.

24 Lage GM, Ugrinowitsch H, Malloy-Diniz LF. Práticas esportivas. In: Malloy-Diniz LF, Fuentes D, Mattos P, Abreu N. Avaliação neuropsicológica. Porto Alegre: Artmed; 2010.
25 Paixão, C.C.; Fortaleza, L.L.; Conte, T. Um estudo preliminar sobre as implicações de tipos de personalidade no ensino de computação. XXXII Congresso da Sociedade Brasileira de Computação (CSBC) — XX Workshop sobre Educação em Informática (WEI), Curitiba–PR; 2012
26 Perez CR, Rabelo IS, Rubio K. Avaliação de traços de personalidade em futuros educadores do esporte brasileiro. Revista Brasileira de Ciência e Movimento. 2013; 21(4): 48-55.
27 Rabelo IS. Investigação de traços de personalidade em atletas brasileiros: análise da adequação de uma ferramenta de avaliação psicológica [tese]. Escola de Educação Física e Esporte, Universidade de São Paulo, São Paulo; 2013.
28 Rabelo IS, Rubio K, Gonçalves GCM, Silva PVC. Monitoring of personality traits among candidates of an athletics program. International Journal of Applied Psychology. 2015; 5(5): 119-125.
29 Andrade BRDR. Transformando suor em ouro. Rio de Janeiro: Sextante; 2006
30 Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. São Francisco: Morgan Kaufman; 2005 [acesso 19 mai. 2018]. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/index.html.
31 Pereira MG, Galvão TF. Etapas de busca e seleção de artigos em revisões sistemáticas da literatura. Epidemiologia e Serviços de Saúde. 2017; 23(2): 369-371.
32 Kitchenham BA. Procedures for Performing Systematic Reviews. Tech. report TR/SE-0401. Newcastle (UK): Keele University; 2004
33 Ferenhof HA, Fernandes RF. Passo-a-passo para construção da Revisão Sistemática e Bibliometria. 2015 [acesso 18 mai. 2018]. Disponível em: http://www.igci.com.br/artigos/passos_rsb.pdf.
34 Nunes F. Palestra "Revisões Sistemáticas" - Profa. Dra. Fátima Nunes (30/09/2015 - MAE/USP). 2015. Disponível em: <https://youtu.be/Wgaw97mTKWM>. Acessado em: 18 mai 2018.
35 Silva EM. Descoberta de Conhecimento com o uso de Text Mining: Cruzando o Abismo de Moore. [dissertação]. Brasília: Universidade Católica de Brasília; 2002
36 Vissoci JRN, Oliveira LPD, Nascimento Junior JRAD, Nakashima FS, Machado WDL, Ciampa ADC, Vieira LF. Esporte é um contexto que possibilita emancipação ou colonização no processo de formação identitária? Revista de psicología del deporte. 2018; 27(4): 59-65.
37 Ramos GP. Processo de Mineração de Desejos aplicado em dados dos Jogos Olímpicos Rio 2016. [trabalho de conclusão de curso]. Rio de Janeiro: Universidade Federal do Estado do Rio de Janeiro; 2016
38 Fengler DI, Frozza R. Classificação de sentimentos e sua validação em comentários de usuários. Anais do Salão de Ensino e de Extensão, 2017, p. 356.
39 Coelho UM, Lima ACE, Omar N. Analisador de expressões positivas e negativas aplicado em comentários de livros e filmes. Simposio Argentino de GRANdes DAtos (AGRANDA)-JAIIO 46, Córdoba, Argentina, 2017.
40 Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and trends in information retrieval. 2008; 2(1-2): 1-135.
41 Cavalcanti DC, Prudêncio RBC, Pradhan SS, Shah JY, Pietrobon RS. Análise de sentimento em citações científicas para definição de fatores de impacto positivo.

Proceedings of the IV International Workshop on Web and Text Intelligence (WTI); 2012. p. 1–10.
42 Liu B. Sentiment analysis and opinion mining. Williston (USA): Morgan & Claypool Publishers; 2012.
43 Guelpeli MVC. Autônomo, Sumarização e Aprendizado [tese]. Rio de Janeiro: Universidade Federal Fluminense; 2008
44 Mantovani TC. Mineração de interesses pessoais a partir de redes sociais para apoiar a personalização de trajetos. [trabalho de conclusão de curso]. Campo Mourão (PR): Universidade Tecnológica Federal do Paraná; 2015.
45 Rubio K, Rabelo IS, Cruz RM. Avaliação de aspectos psicológicos em Educação Física e Esporte. In: Böhme, MTS (org). Avaliação e Desempenho em Educação Física e Esporte. São Paulo: Editora Manole; 2018
46 Nunes, MASN. Computação Afetiva personalizando interfaces, interações e recomendações de produtos, serviços e pessoas em Ambientes computacionais. In: Ordonez NO (org.). São Cristóvão: DCOMP/PROCC/UFS; 2012. p. 115–151.
47 Peirce CS. The Collected Papers of Charles Sanders Peirce. Cambridge, Massachusetts: Harvard University Press; 2005.