

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

MARCELO VICENTE BERALDI

ROBÔS DE INVESTIMENTO A PARTIR DE DADOS DE REDES SOCIAIS

SÃO PAULO
2020

MARCELO VICENTE BERALDI

ROBÔS DE INVESTIMENTO A PARTIR DE DADOS DE REDES SOCIAIS

Dissertação apresentada à Escola de Economia de São Paulo da Fundação Getulio Vargas, como requisito para obtenção do título de Mestre em Economia.

Area de Concentração: Finanças.

Orientador: Prof. Dr. Marcelo Fernandes

SÃO PAULO
2020

Beraldi, Marcelo Vicente.

Robôs de investimento a partir de dados de redes sociais / Marcelo Vicente Beraldi.
- 2020.

31 f.

Orientador: Marcelo Fernandes.

Dissertação (mestrado profissional MPFE) – Fundação Getulio Vargas, Escola de
Economia de São Paulo.

1. Ações (Finanças). 2. Investimentos. 3. Inteligência artificial. 4. Sistemas
especialistas (Computação). I. Fernandes, Marcelo. II. Dissertação (mestrado
profissional MPFE) – Escola de Economia de São Paulo. III. Fundação Getulio Vargas.
IV. Título.

CDU 336.767

MARCELO VICENTE BERALDI

ROBÔS DE INVESTIMENTO A PARTIR DE DADOS DE REDES SOCIAIS

Dissertação apresentada à Escola de Economia de São Paulo da Fundação Getulio Vargas, como requisito para obtenção do título de Mestre em Economia.

Area de Concentração: Finanças.

Data de aprovação

25 /11/ 2020

Banca examinadora:

Prof. Dr. Marcelo Fernandes (orientador)
FGV EESP

Prof. Dr. Fernando Daniel Chague
FGV EESP

Prof. Dr. Hellinton Takada (Co-Orientador)
Santander Brasil

RESUMO

Com a difusão das opções de investimentos e a busca por experiência diferenciada ao cliente, muitos bancos e corretoras começaram a desenvolver consultores automatizados de investimento (mais conhecidos como *Robo-Advisors*) para fornecer ao cliente um portfólio personalizado sem a interação de um agente humano. Contudo, sua difusão se mostrou um desafio em vários aspectos, em especial pelas limitações dos modelos tradicionais de media-variância, prejudicando seu uso ao ponto de muitos bancos e corretoras abortarem seus projetos. O objetivo desse trabalho é sugerir uma abordagem alternativa de construção de carteira personalizada a partir de dados de redes sociais em conjunto com técnicas difundidas na indústria de investimentos. Para tal, usam-se métodos de aprendizado de máquina a fim de prover elementos em tempo real para maximizar o retorno das carteiras.

Palavras-chave: *Fintechs*, *Robo advisors*, Redes sociais, Inferência, *Big data*, Finanças, *Machine Learning*, Alocação de carteiras, Aprendizado de Máquina, Automação, *Risk Parity*, *Black&Litterman*, Paridade de risco

ABSTRACT

Considering the recent spread of investment products and the constant search for a differentiated customer experience, many banks and brokers are developing automated investment advisors (also called *Robo-Advisors*) which aims to provide a personalized portfolio without a direct human interaction. Although *Robot-Advisors* diffusion proves to be a challenge in several aspects, especially due to the limitations of traditional models of mean-variance, hampering its use to the point that many banks and brokers have aborted such projects. The scope of this paper is suggesting an alternative approach to building a personalized portfolio from social network data as resource for most used allocation theories. For this, machine learning methods are used in order to provide elements in real time to maximize the return of the portfolios.

Keywords: Fintechs, Robo advisors, Finance, Social networks, Inference, Big data applied in Finance, Machine learning applied in finance, Portfolio allocation, Machine learning, automation, Risk Parity, Black & Litterman,

LISTA DE FIGURAS

Figura 1 – Pesquisa de Beketov, Lehmann, e Wittke (2018)	11
Figura 2 – Gráfico de retornos acumulados, sem pesos de redes sociais.....	24
Figura 3 - Alocação média de ativos no modelo de BL sem redes sociais	25
Figura 4 - Alocação média de ativos no modelo de PR sem redes sociais	25
Figura 5 – Indicadores com redes sociais.....	26
Figura 6 – Média ponderada Alocação Ativos Black&Litterman com redes sociais e Alpha de 15%....	27
Figura 7 – Média ponderada Alocação Ativos Black&Litterman com redes sociais e Alpha de 30%....	27
Figura 8 – Média ponderada Alocação Ativos Risk Parity com redes sociais e Alpha de 15%.....	28
Figura 9 – Média ponderada Alocação Ativos Risk Parity com redes sociais e Alpha de 30%.....	28

LISTA DE TABELAS

Tabela 1– Índices B3 e sua Composição	22
Tabela 2– Indicadores sem redes sociais	26
Tabela 3 – Indicadores com redes sociais	28

SUMÁRIO

1. INTRODUÇÃO	10
2. REVISÃO DA LITERATURA	11
3. METODOLOGIA.....	13
3.1. Bases de Dados.....	21
4. RESULTADOS	23
4.1. Robô de investimento sem inferência de redes sociais	24
4.2. Robô de investimento com inferência de redes sociais.....	26
5. CONCLUSÕES.....	29
6. REFERÊNCIAS.....	30

1. INTRODUÇÃO

O mercado de investimentos focados em pessoa físicas (alta renda e varejo) cresceu 13% no Brasil em 2019, totalizando um mercado de 3,3 trilhões de reais em ativos. Em particular, observou-se um crescimento significativo nos ativos considerados de maior risco, em razão principalmente da diminuição da taxa básica de juros brasileira, como mostra o relatório anual da ANBIMA (2019).

Esse mercado em rápida transformação tem favorecido o surgimento e consolidação de empresas de tecnologias financeiras (conhecidas por *Fintechs*) propondo métodos inovadores de gestão de portfólio de investimento. Um dos modelos mais utilizados consiste no uso de robôs de consultoria de investimento (ou *robo-advisors*, em inglês) para propor uma carteira personalizada ao cliente. Em geral, os robôs utilizam métodos quantitativos convencionais, tendo como parâmetros questionários padronizados que os clientes respondem para identificar seu perfil de risco. Além disso, contam com parâmetros pré-definidos pelos gestores que adicionam sua visão em relação ao mercado.

Essa abordagem apresenta várias limitações (FEIN, 2015) a ponto de as autoridades reguladoras emitirem comunicados alertando para os riscos potenciais (FINRA, 2016). Em linhas gerais a autoridade financeira americana defende que o modelos de otimização e diversificação de carteira usados pelos robôs de investimento não são eficientes, sugerindo que principais decisões de alocação seja tomadas por gestores humanos. Em outras palavras, na prática, os atuais “*robo-advisors*” automatizam apenas algumas tarefas inerentes à assessoria convencional, proporcionando um conjunto de soluções equivalentes aceitáveis que precisam ser posteriormente filtradas pelo assessor humano.

Contudo, existe uma demanda crescente por gestão de ativos através de *robo-advisors*, com mínima interação direta de elemento humano. A minimização da interação humana também traz algumas vantagens, como o menor conflito de interesses e, possivelmente, maior eficiência de mercado (BEKETOV, LEHMANN, e WITTKE, 2018). Pesquisas recentes apontam que essa indústria vem crescendo vertiginosamente e deve chegar a movimentar 1,1 trilhões de dólares em ativos até o final de 2020 (STATISTA 2020).

Esse trabalho visa contribuir para a literatura de robôs de investimento, fornecendo maneiras de depurar o conjunto de carteiras personalizadas aceitáveis no mercado brasileiro. Para tal, usamos ferramentas robustas de aprendizado profundo por máquina a fim de pontuar

e ordenar as soluções advindas dos robôs tradicionais. Em particular, simulamos possíveis saídas de um robô tradicional que utilizamos como entrada para um protótipo mais avançado de “*robo-advisor*”.

O restante da dissertação estrutura-se da seguinte forma. O capítulo 2 faz uma Revisão Da Literatura. O capítulo 3 descreve o Modelo, discutindo detalhadamente sua origem, características e objetivos. O capítulo 4 discorre sobre a Metodologia empregada na busca pela carteira personalizada. O capítulo 5 documenta os Resultados encontrados. Por fim, o capítulo 6 conclui.

2. REVISÃO DA LITERATURA

Como argumentado por Beketov, Lehmann, e Wittke (2018) as teorias que sustentam as decisões de alocação de recursos por robôs de investimento não são novas, sendo a *Modern Portfolio Theory* a mais relevante como vemos na figura 1. Postulada primeiramente por Markowitz (1952), sua abordagem de média-variância serve de base para a maioria dos robôs de investimento que depois recebem parâmetros específicos de forma a delinear a estratégia de alocação proposta por seus gestores.

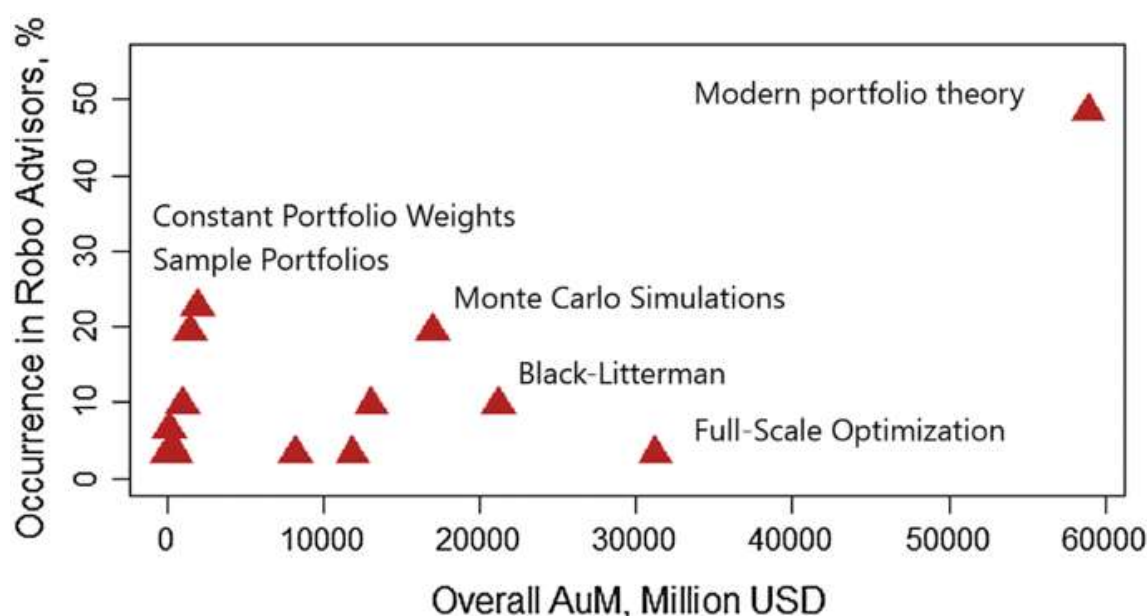


Figura 1 – Pesquisa de Beketov, Lehmann, e Wittke (2018) com gestores de fundos que usam Robo Advisors

Contudo, a ausência de interferência humana, em especial no rebalanceamento da carteira, leva a uma série de implicações e riscos (FINRA 2016), em especial, na compreensão do perfil de risco do investidor e na gestão de risco. Tais ressalvas são em grande parte fruto do modelo matemático que suporta a alocação de ativos baseada em média-variância (MV). Embora esse método, proposto por Markowitz (1952), tenha apoiado por décadas as decisões de gestores, existem limitações conhecidas que tornam os resultados economicamente pouco intuitivos. Michaud (1989) discorre sobre distorções causadas pela inversão da matriz de covariância quando o número de ativos é grande. Na prática, o processo de otimização maximiza o erro de apreçamento em vez da compensação pelo risco, resultando em carteiras altamente concentradas em alguns poucos ativos e, portanto, pouco diversificadas. Outro ponto se refere a falta de estabilidade da carteira. As mudanças bruscas na composição da carteira a cada rebalanceamento prejudica retorno da mesma por conta dos custos mais altos de transação.

Diversos trabalhos abordam tais características e apontam técnicas para melhorar os estimadores. Black e Litterman (1990) propõem um modelo que incorpora a visão dos gestores. Esse modelo utiliza-se de uma distribuição de mercado (ou priori) calculada usando o método tradicional MV e a visão de mercado do gestor combinadas pelo teorema de Bayes, resultando numa distribuição posteriori. Essa combinação mitiga muitos dos problemas de erro de estimação contudo outros problemas como soluções de canto e concentração em ativos permanecem. O modelo comporta uma visão bastante completa a fim de garantir a diversificação contudo a posteriori se distancia muito da priori, comprometendo a qualidade do estimador resultante.

Nesse cenário, Roncalli (2013) propõe um modelo de paridade de riscos (RP), que considera a contribuição de risco de cada ativo na composição da carteira. Cada ativo contribui assim de forma equivalente para o risco da carteira, resultando em uma carteira mais diversificada uma vez que todos os ativos recebem alguma alocação. Se as volatilidades dos ativos não mudarem de maneira significativa em janelas curtas de rebalanceamento, a composição da carteira é mais estável ao longo do tempo e, portanto, mais viável na prática. Bourgeron, Lezmi, e Roncalli (2018) incorporam técnicas de regularização a fim de tornar menos volátil a alocação, das quais podemos destacar Tikhonov (Groetsch 1984), Ridge (Hoerl e Kennard 1970) e Lasso (Tibshirani 1996). Tais regularizações são o “ingrediente secreto” para que os *Robo advisors* tomem vida e sejam eficientes. Para o mercado brasileiro, não temos um número grande de ativos candidatos, portanto deixamos o uso dessas técnicas como sugestão para trabalhos futuros.

Os modelos *Black&Litterman* e *Risk Parity* consideram as premissas de seus gestores. Neste trabalho, sugerimos ferramentas de aprendizado (profundo) por máquina para delimitar essas premissas a partir da enorme massa de dados disponíveis em plataformas de notícias e redes sociais. Para tal, empregamos análises de sentimento, como o *VADER (Valence Aware Dictionary for sEntiment Reasoning)* (Hutto e Gilbert 2014) no qual utiliza-se um dicionário para inferir se um comentário é positivo ou negativo, possibilitando analisar uma determinada opinião autonomamente, sem o olhar humano. Com essa ferramenta podemos monitorar diversos canais de notícias e redes sociais, buscando um determinado ativo, segmento ou grupo de empresas a fim de gerar um conjunto de sentimentos quantificáveis que possam ser transformados em entrada para o nosso modelo.

3. METODOLOGIA

Os robôs de investimento utilizam vários métodos de otimização de portfólio, com inúmeras estratégias de balanceamento. O procedimento de paridade de risco, desenvolvido por *Roncalli* (2013), está entre as estratégias de otimização mais populares. Inicia-se com um programa de otimização de média-variância à la *Markowitz* (1952):

$$x^*(\gamma) = \arg \min \frac{1}{2} s^T \sum x - \gamma x^T (\gamma - r1) \quad (1)$$

em que x^* denota o peso ótimo do portfólio, γ a aversão ao risco do investidor e r a taxa livre de risco. No entanto, *Bourgeron, Lezmi, e Roncalli* (2018) argumentam que a solução x^* de *Markowitz* não representa um ponto ótimo financeiro, seguindo o raciocínio de *Michaud* (1989). *Roncalli* (2014) propõe uma abordagem diferente a fim de aumentar a diversificação. Primeiro, reescrevemos a equação objetivo para minimizar o risco:

$$x^*(c) = \arg \min \mathcal{R}(x), \quad (2)$$

em que:

$$\mathcal{R}(x) = -\pi(x) + c. \sigma(x), \quad (3)$$

$\pi(x) = \gamma(x) - r1$ é o prêmio de risco da carteira com peso x e $\sigma(x) = \sqrt{x^T \Sigma x}$ é a volatilidade da carteira.

Podemos assim escrever a contribuição de risco de cada ativo i como

$$\mathcal{RC}_i = -\pi_i(x) + c \cdot \sigma_i(x) \quad (4)$$

e a contribuição normalizada por:

$$\mathcal{RC}_i^* = \frac{-\pi_i(x) + c \cdot \sigma_i(x)}{R(x)}. \quad (5)$$

Analogamente, definimos o excesso de retorno normalizado PC_i^* e a volatilidade normalizada VC_i^* como

$$PC_i^* = \frac{-\pi_i}{-\pi} = \frac{x_i \cdot \pi_i}{\sum_{j=1}^n x_j \cdot \pi_j} \quad (6)$$

$$VC_i^* = \frac{\sigma_i(x)}{\sigma(x)} = \frac{x_i \cdot (\Sigma x)_i}{x^T \Sigma x} \quad (7)$$

Portanto, podemos reescrever o programa de otimização como um problema geral de *Risk Budgeting* (RB):

$$\mathcal{RC}_i^*(x) = b_i \mathcal{R}(x) \quad (8)$$

em que b_i corresponde à quota de risco do ativo i expressa em termos relativos ao risco total $\mathcal{R}(x)$. Ademais, $b_i > 0$ é uma condição necessária para a estabilidade do modelo. Lembrando que b_i e x_i são quotas de risco e peso do ativo na carteira para o ativo i portanto seus valores são sempre positivos e a $\sum_{i=1}^n x_i = 1$ e $\sum_{i=1}^n b_i = 1$, isso é, não permite venda a descoberto e alavancagem, sendo n o número de ativos.

Roncalli (2013) demonstra que numa situação econômica real na qual temos alocação em todos os ativos ($b_i > 0$) e não temos alavancagem (isso é, o valor do ativo, mesmo que o retorno seja $\mathcal{RC}_i - \infty$, é maior ou igual a zero), a solução converge e é única.

Vários autores propuseram soluções computacionais para encontrar a solução x^* como *Chaves et al.* (2012), *Griveau-Billion, Richard, e Roncalli* (2013). Para esse trabalho propomos a solução numérica através do SCRIP (Feng e Palomar, 2015) pois apresenta uma eficiência computacional maior quando sujeita a várias restrições, característica que interessa para comportar inferências de redes sociais como veremos mais adiante. Mais especificamente, *Feng e Palomar* (2015) propõem minimizar

$$\begin{aligned} U^*(w) &\triangleq \mathcal{R}(w) + \lambda F(w) \\ \text{sujeito a } w^T 1 &= 1 \text{ e } w \in W \end{aligned} \quad (9)$$

sendo $\mathcal{R}(w)$ a concentração de risco definida por uma função quadrática:

$$\mathcal{R}(w) \triangleq \sum_{i=1}^n (g_i(w))^2 \quad (10)$$

em que w são os pesos dos ativos no portfólio, $g_i(w)$ uma função diferenciável não convexa que retorna a concentração de risco do i ativo, $F(w)$ é uma função convexa que representa as opções do portfólio, λ é um parâmetro maior que zero que balanceia concentração do portfólio, $w^T 1 = 1$ é a restrição de capital e W um conjunto convexo de premissas, que no nosso modelo serão os pesos dos insights das redes sociais.

Como solução computacional, Feng e Palomar (2015) convexificam $\mathcal{R}(w)$ através da linearização de cada termo $g_i(w)$ interno ao operador quadrático da somatória. No mais, adicionam o termo $\|w - w^k\|_2^2$ para garantir convergência, obtendo

$$\begin{aligned} \min (w) \quad & P(w; w^k) + \frac{\|w - w^k\|_2^2}{2} + \lambda F(w) \\ \text{sujeito a } & w^T 1 = 1 \text{ e } w \in W \end{aligned} \quad (11)$$

em que

$$P(w; w^k) \triangleq \sum_{i=1}^n \left(g_i(w^k) + (\nabla g_i(w^k))^T (w - w^k) \right). \quad (12)$$

$P(w; w^k)$ possui o mesmo gradiente de $\mathcal{R}(w)$ portanto converge para o mesmo ponto ótimo. Além disso, trata-se de uma função quadrática computacionalmente leve. Podemos ainda reescrever $P(w; w^k)$ em função dos seus gradientes como:

$$P(w; w^k) = \|A^k(w - w^k) + g(w^k)\|_2^2 \quad (13)$$

em que

$$\begin{aligned} A^k &\triangleq [\nabla g_1(w^k), \nabla g_2(w^k), \dots, \nabla g_i(w^k)]^T \quad e \\ g(w^k) &\triangleq [g_1(w^k), g_2(w^k), \dots, g_i(w^k)]^T. \end{aligned} \quad (14)$$

Por fim, podemos escrever o problema em termos computacionais como:

$$\begin{aligned} \min (w) \quad & \frac{w^T Q^k w}{2} + \frac{w^T q^k}{2} + \lambda F(w) \\ \text{sujeito a } & w^T 1 = 1 \text{ e } w \in W \end{aligned} \quad (15)$$

sendo

$$Q^k \triangleq 2(A^k)^T A^k + \lambda I \quad e \quad q^k \triangleq 2(A^k)^T g(w^k) - Q^k w^k \quad (16)$$

Se considerarmos que W possui apenas restrições lineares podemos escrever $w^T 1 = 1$ como $Cw = c$. Sendo C uma matriz com restrições não colineares podemos assumir que todas as condições de Karush-Kuhn-Tucker (KKT) são válidas e reescrever o problema acima como:

$$\hat{w}^k = -(Q^k)^{-1}(q^k + C^T \lambda^k) \quad (17)$$

em que

$$\lambda^k = -\left(c(q^k)^{-1}c^T\right)^{-1}\left(c(q^k)^{-1}q^k + c\right), \quad (18)$$

sendo $w^{k+1} = w^k + \gamma(\hat{w}^k - w^k)$ a próxima interação até a convergência.

Nesse ponto vale descrever com mais detalhe a matriz C e o escalar λ . Seus valores são arbitrários, isso é, no modelo são geralmente premissas dos investidores. No modelo que propomos, alimentamos tais premissas usando informações dispersas nas redes sociais e de notícias. C representa um conjunto de restrições lineares. Um exemplo para $i=8$ (ativos) como descrito por (Fast Design of Risk Parity Portfolios [s.d.]) poderia ser:

$$\begin{cases} w_5 + w_6 + w_7 + w_8 \geq +30\% \\ w_2 + w_6 \geq w_1 + w_1 + 5\% \end{cases} \quad (19)$$

que resulta numa matriz C igual a

0	0	0	0	-1	-1	-1	-1
1	-1	0	0	1	-1	0	0

(20)

e matriz c igual a

-0,30
-0,05

(21)

Imaginamos agora um exemplo para o escalar λ . Como descrito na equação (9), quanto mais próximo de zero, menor o peso de $F(w)$, reduzindo a importância das premissas em C . No limite se $\lambda = 0$, os riscos são distribuídos uniformemente, sem considerar as premissas impostas pelo aprendizado por máquina.

Adicionamos o fundamento teórico sobre interpretação de texto de redes sociais/redes de notícias. Esse tema é pesquisado em várias áreas do conhecimento, desde medicina a antropologia, incluindo vários estudos no campo das finanças. Podemos destacar *Garcia e Schweitzer* (2015) e *Kim* (2016) que documentam correlação entre os preços dos ativos e o sentimento social. Na mesma direção aponta o estudo de *Ke, Kelly, e Xiu* (2019) que encontra forte correlação entre os sentimentos de 30 milhões de textos do microblog *Twitter* e

a direção dos retornos de índices de ativos. Pode ainda adicionar o trabalho de *Nguyen, Shirai, e Velcin* (2015) demonstrando que o sentimento humano fornece um dos maiores poderes explicativos à movimentação de ativos acionários comparado a outras técnicas de interpretação de texto.

Sendo assim a quantificação do sentimento humano a partir de textos de redes sociais é tem o potencial de fornecer elementos indicativos para melhorar a previsão dos retornos. Em um campo com diversas técnicas como destacam os trabalhos de *Hutto e Gilbert* (2014) e *Lei, Qian, e Zhao* (2016), podemos indicar que a técnica *Vader* (*Valence Aware Dictionary for sEntiment Reasoning*) fornece um pode interpretação de texto de *microblogs* maior comparados a outras técnicas como *SVM* e *Naive Bayes* o que a torna uma boa candidata ao modelo a ser proposto. As técnicas de mensuração de sentimento social consistem em confrontar sentenças, conhecidas também pelo termo *Corpus*, com bases de treinamento previamente avaliadas, conhecidas como dicionários, a fim de quantificar numericamente o sentimento contido naquele *Corpus*. Uma sentença simples como “Eu estou feliz hoje” é facilmente interpretada como positiva por uma pessoa. Na proposta de replicar tal percepção as técnicas de análise de sentimento buscam palavras individuais ou em sequência que exprimem tal sentimento. No exemplo do dicionário utilizado pela técnica *VADER* a resposta seria {0, 0, 2.7, 0}, isso é, a palavra “feliz” possuiu um sentimento positivo de 2.7 enquanto as outras palavras não tem sentimento.

Alterando levemente a frase para “Estou muito feliz hoje”, temos um resultado semelhante uma vez que a palavra introduzida não tem sentimento associado quando vista isoladamente. Contudo se a colocamos junto com uma palavra com sentimento ela a modifica, aumentando ou diminuindo a sua pontuação. Nesse exemplo podemos intuir que “muito feliz” tem um sentimento maior que a palavra isolada “feliz”. Na técnica *VADER*, que considera sequências de até três palavras, a pontuação seria {0, 0, 0, 2.993, 0}.

Importante salientar que o dicionário é fruto de uma base de treinamento previamente avaliada por humanos. Dessa maneira a qualidade do dicionário se torna tão importante quanto a técnica em si. Quanto maior o número de sequências de palavras e quanto maior o escrutínio dessas sequências por mais avaliadores, mais confiável será a pontuação.

As técnicas mais difundidas como citadas por *Hutto e Gilbert* (2014) pode ser divididas em dois grandes grupos: semânticos em que a resposta só pode ter valores negativo, positivo

ou neutro; e múltiplas valências em que uma sequência de palavras tem valor real em um determinado intervalo.

As técnicas semânticas são especialmente úteis quando o interesse do indicador é positivo ou negativo, isso é, se a pergunta de pesquisa é sim ou não. Tal técnica seria útil por exemplo se um administrador público propõe uma política e deseja interpretar os comentários de uma publicação nas redes sociais e saber a proporção de seus eleitores que são contra ou a favor de tal proposta. No entanto, se a pergunta de pesquisa é sobre a intensidade das opiniões, as técnicas de múltiplas valências são mais informativas. A metodologia proposta nesse trabalho visa utilizar um índice de sentimento como premissas dos gestores nos modelos de alocação de ativos *BL* e *RP*. Sendo tais premissas números reais com intervalo de zero a 1 (hum) um modelo de valência fornece a saída apropriada.

Dentre as técnicas de múltiplas valências devemos selecionar aquela que é mais apropriada ao dado que se deseja interpretar. Esse trabalho emprega exclusivamente dados de *Twitter*, que possuem um formato característico pela limitação de caracteres (120 ao máximo, excluindo marcações e links) e pela informalidade, levando ao uso excessivo de pontuações, espaços e símbolos de emoções. Com essa motivação, Hutto e Gilbert (2014) adaptam alguns conceitos de múltiplas valências no *VADER* para ter especial aderência a textos de redes sociais.

Vale destacar que, assim como a maioria das técnicas de múltiplas valências, temos dicionários predominantemente em inglês, restando como alternativa mais viável a tradução dos textos, mantendo a pontuação e símbolos, e aplicando o método aos dados traduzidos. Embora com isso adiciona-se um erro de um possível erro de tradução, Mohammad, Salameh, e Kiritchenko (2016) e Elnagar, Einea, e Lulu (2017) defendem que o sentimento não se altera de maneira substancial podendo ser usado em um grande número de dados. Ao final teremos um conjunto de comentários num intervalo de tempo com seus respectivos sentimentos. Vale a pena destacar que o número absoluto (ou valência) desse sentimento é pouco importante, estamos mais interessados na variação do sentimento que nos dará uma magnitude da variação da alocação de tal ativo.

A análise de sentimentos pode não dar informações sobre a melhor alocação de ativos, contudo indica bem o sentido de movimentação do mercado no dia seguinte (LEI, QIAN, e ZHAO (2016). Sendo assim podemos quantificar o impacto do sentimento como:

$$w'_{d+1} = w'_d \cdot \left(\frac{S_d}{\sum S} \right) \cdot \psi \quad (22)$$

em que w'_d é o vetor de pesos do dia d , $\left(\frac{S_d}{\sum S} \right)$ uma matriz da diferença entre as mediana dos sentimentos dos ativos w'_d no dia d e de todos os sentimentos S (ou sentimento neutro) e ψ um escalar que ajusta os pesos de forma que $\sum w'_{d+1} = 1$.

Definindo $\left(\frac{S_d}{\sum S} \right)$ uma matriz diagonal com números positivos, em que cada elemento i da diagonal é um multiplicador de cada peso $w'_d i$ de modo a aumentar ou diminuir sua alocação no dia seguinte $w'_{d+1} i$. Essa matriz $\left(\frac{S_d}{\sum S} \right)$ pode ser definida de várias formas o que a assemelha bastante a opção típica de premissas de gestores, um traço comum entre diferentes métodos de alocação de ativos. Podemos então definir cada elemento como:

$$S_d^i = \left(\frac{(S_d^{ii} - S_{d \min}^i)}{(S_{d \max}^i - S_{d \min}^i)} \right) \alpha + \beta, \quad (23)$$

sendo S_d^{ii} o sentimento do ativo i no dia d , $S_{d \min}^i$ o valor mínimo dos sentimentos dentre todos os ativos no dia d , $S_{d \max}^i$ o valor máximo dos sentimentos dentre todos os ativos no dia, α uma variação máxima percentual da alocação entre o menor e maior ativo e β o percentual mínimo de alocação do ativo de menor valor. Dessa maneira, os elementos S_d^i são porcentagens de variação dos pesos dos ativos i dentro do intervalo $[\beta, \alpha + \beta]$. Já w'_{d+1} pode ser escrito como

$$w'_{d+1} = [w_d^1 S_d^1, w_d^2 S_d^2, \dots, w_d^n S_d^n] \quad (24)$$

em que n é o número de ativos e $w_d^n S_d^n$ é o peso mínimo desse ativo na carteira. Sendo assim w'_{d+1} tem o formato compatível com as restrições lineares do método do modelo *SCRIP* (Feng e Palomar 2015) atendendo o objetivo de fornecer um modelo que integra métodos de alocação de ativos estocásticos com sentimentos difusos em mídias sociais.

No caso do modelo de BL a matriz $\left(\frac{S_d}{\sum S} \right)$ é mais simples, uma vez que se admitem premissas relativas, isso é, podemos explicitar que ativos têm maior probabilidade de se valorizarem em relação a outros pelo seu sentimento social. Nesse caso, temos cada elemento definido como

$$S_d^i = \frac{S_d^{ii}}{\left(\frac{\sum_{i=1}^n S_d^{ii}}{n}\right)} \alpha \quad (25)$$

e, portanto,

$$w'_{d+1} = [S_d^1, S_d^2, \dots, S_d^n] \quad (26)$$

com cada elemento S_d^n fornecendo um peso relativo como descreve o método BL sobre a incorporação da visão dos gestores.

3.1. Bases de Dados

Como descrito por Bourgeron, Lezmi, e Roncalli (2018), classes de ativos seriam mais recomendadas quanto comparados a tipos de ativos pois apresentam autovetores mais significativos e permitem que as técnicas de seleção de portfolio obtenham resultados mais significativos e que o risco idiossincrático seja menor. Essa é uma das motivações que leva a maioria dos robôs de investimento a utilizarem apenas *ETFs* (Exchange Traded Funds), isso é, fundos de ações negociados em bolsa. Diminui-se assim o número de ativos analisados, permitindo uma maior robustez dos métodos quantitativos e contribuindo para aumentar a diversificação ao mesmo tempo que contribui para uma significativa redução nos custos de transação (Kaya 2017).

O mercado de ETF brasileiro é ainda bastante incipiente, contando com apenas 17 fundos listados na B3, replicando 4 índices. Para título de comparação, há mais de 2.000 ETFs nos EUA e na Europa, totalizando mais de 5.000 globalmente (ETFGI - Independent ETFs / ETPs Research and Consultancy Firm [s.d.]). A solução encontrada foi usar os índices de segmento da B3, num total de 7 índices (Índices de Segmento | B3 [s.d.]), que serão considerados “ativos” embora nem todos tenham ETFs seguindo sua composição. O uso de índices tem algumas vantagens, com a boa disponibilidade histórica e ajustes tanto na composição quando monetária. Contudo torna o exercício meramente teórico, pelo menos enquanto o mercado de *ETFs* brasileiro for pouco robusto. Usamos o IBOVESPA como ativo referência. O objetivo é obter retornos acima do IBOVESPA, portanto.

Índice B3	Ativos que compõem o índice (código B3)
ICON – Consumo	ALPA4, ABEV3, ANIM3, ARZZ3, BTOW3, BKBR3, BRFS3, CAML3, CRFB3, CEAB3, CNTO3, HGTX3, COGN3, CVCB3, CYRE3, DIRR3, EVEN3, EZTC3, FLRY3, GFSA3, GRND3, NTCO3, GUAR3, HAPV3, HBOR3, HYPE3, PARD3, MEAL3, GNDI3, MYPK3, JBSS3, JHSF3, RENT3, LCAM3, LAME3, LAME4, AMAR3, LREN3, MDIA3, MGLU3, MRFG3, LEVE3, BEEF3, MOVI3, MRVE3, ODPV3, PCAR3, QUAL3, RADL3, SMTO3, SEER3, SLCE3, SMLS3, TCSA3, TEND3, TRIS3, VVAR3, VIVA3, VULC3
IEEX – Energia Elétrica	TIET11, ALUP11, CMIG4, CESP6, COCE5, CPLE6, CPFE3, CPRE3, ELET3, ENBR3, ENGI11, ENEV3, EGIE3, EQTL3, LIGT3, NEOE3, OMGE3, TAEE11, TRPL4
IFNC – Financeiro	ABCB4, B3SA3, BIDI4, BIDI11, BPAN4, BRSR6, BBSE3, BBDC3, BBDC4, BBAS3, BPAC11, CIEL3, IRBR3, ITSA4, ITUB3, ITUB4, PSSA3, SANB11, SULA11
IMAT – Materiais básicos	BRAP4, BRKM5, DTEX3, GGBR4, GOAU4, KLBN11, CSNA3, SUZB3, UNIP6, USIM5, VALE3
IMOB – Imobiliário	ALSO3, BRML3, BRPR3, CYRE3, DIRR3, EVEN3, EZTC3, GFSA3, HBOR3, IGTA3, JHSF3, LOGG3, MRVE3, MULT3, TCSA3, TEND3, TRIS3
INDX – Setor Industrial	ALPA4, ABEV3, BRKM5, BRFS3, CAML3, HGTX3, CYRE3, DIRR3, DTEX3, EMBR3, EVEN3, EZTC3, GFSA3, GGBR4, GOAU4, GRND3, NTCO3, HBOR3, MYPK3, JBSS3, JHSF3, KLBN11, MDIA3, POMO4, MRFG3, LEVE3, BEEF3, MRVE3, POSI3, RAPT4, SMTO3, CSNA3, SUZB3, TCSA3, TEND3, TRIS3, TUPY3, USIM5, VIVA3, WEGE3
UTIL – Utilidade Pública	TIET11, ALUP11, CMIG3, CMIG4, CESP6, CSMG3, CPLE3, CPLE6, CPFE3, ELET3, ELET6, ENBR3, ENGI11, ENEV3, EGIE3, EQTL3, LIGT3, NEOE3, OMGE3, SBSP3, SAPR4, SAPR11, TAEE11

Tabela 1– Índices B3 e sua Composição

Já nas redes sociais existe um desafio adicional visto que as principais plataformas vem restringindo a capacidade de busca em suas bases, em especial pelos recentes usos impróprios de suas informações a fim de interferir nas eleições americanas (Cadwalladr e Graham-Harrison, 2018). Portanto a base será composta de textos do microblog *Twitter*TM onde buscaremos nas janelas de balanceamento opiniões de usuários sobre os ativos, identificados pelo seus códigos de ativo da B3 e depois o agruparemos por índice de segmento.

Por exemplo, o índice do setor financeiro inclui a ação do Itaú Holding, identificado pelo código B3 *ITUB4*. A busca se dá pelo nome desse ativo o que limita consideravelmente os comentários a opiniões de analistas e investidores, eliminando assim opiniões sobre a empresa em si e seus clientes em suas atividades. Os dados são então traduzidos e sentimentalizados através da técnica *VADER*, obtendo-se assim o sentimento dos investidores e analistas do fechamento do mercado do dia anterior ao fechamento do mercado no dia analisado.

4. RESULTADOS

Os resultados foram obtidos através da análise dos dados de fechamento dos índices setoriais entre 01/01/2019 até 04/06/2020. A janela temporal de aprendizado será de 50% para aprendizado, tendo os outros 50% como conjunto de prova, sendo que a janela de aprendizado se move tornando o histórico cumulativo. Portanto, temos dados de sentimento social de 17/09/2019 até 04/06/2020, somente nas datas de rebalanceamento, isso é, a cada 4 pregões. Infelizmente, não temos uma grande flexibilidade uma vez que a base de dados do microblog *Twitter*TM não permite grande gama de pesquisas. Analisamos os sentimentos do fechamento do pregão anterior até o fechamento do pregão corrente, totalizando 38 mil textos, dos quais identificamos sentimento em 18 mil, numa média de 500 *corpus* por janela de balanceamento. Podemos dizer que é uma base pequena visto que por exemplo *Ke, Kelly, e Xiu* (2019) utilizam 32 milhões de textos.

São confrontados 2 métodos de seleção de portfólio: *Black&Litterman* (BL) e paridade de risco (RP), sempre confrontando seus retornos com o IBOV no mesmo período. Primeiramente analisamos os métodos sem premissas e depois com o sentimento social, sempre com balanceamento a cada 4 pregões. Custos de transação e impostos serão descartados para simplificar a implementação.

4.1. Robô de investimento sem inferência de redes sociais

Nesta seção, apresentamos o desempenho dos modelos de BL e PR sem nenhuma premissa ou inferência, nem mesmo arbitral. Ambos operaram, portanto, sem restrições lineares, buscando-se apenas a tangente da curva de risco-retorno e o portfólio que equaliza as contribuições de risco dos ativos. No caso do BL não é incluída uma *priori*, simplificando o modelo a média-variância.

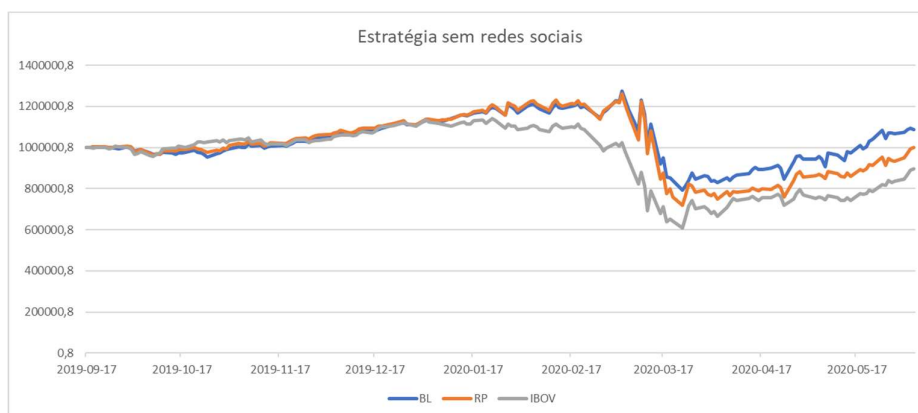


Figura 2 – Gráfico de retornos acumulados, sem pesos de redes sociais

Podemos ver na Figura 1 que o BL apresenta um retorno final maior em relação ao RP e ao IBOV. Os índices de Sharpe de ambos os métodos são levemente superiores àquele do IBOV, como seria esperado já que a função objetivo desses métodos não visa diminuir exclusivamente a volatilidade e mas como também aumentar o retorno.

Uma enorme vantagem do modelo RP fica clara na comparação da composição do portfólio. Vemos que o RP manteve uma alocação bem mais balanceada, diminuindo a exposição de risco a setores específicos, enquanto o modelo BL depende muito do desempenho do ativo IEEX.

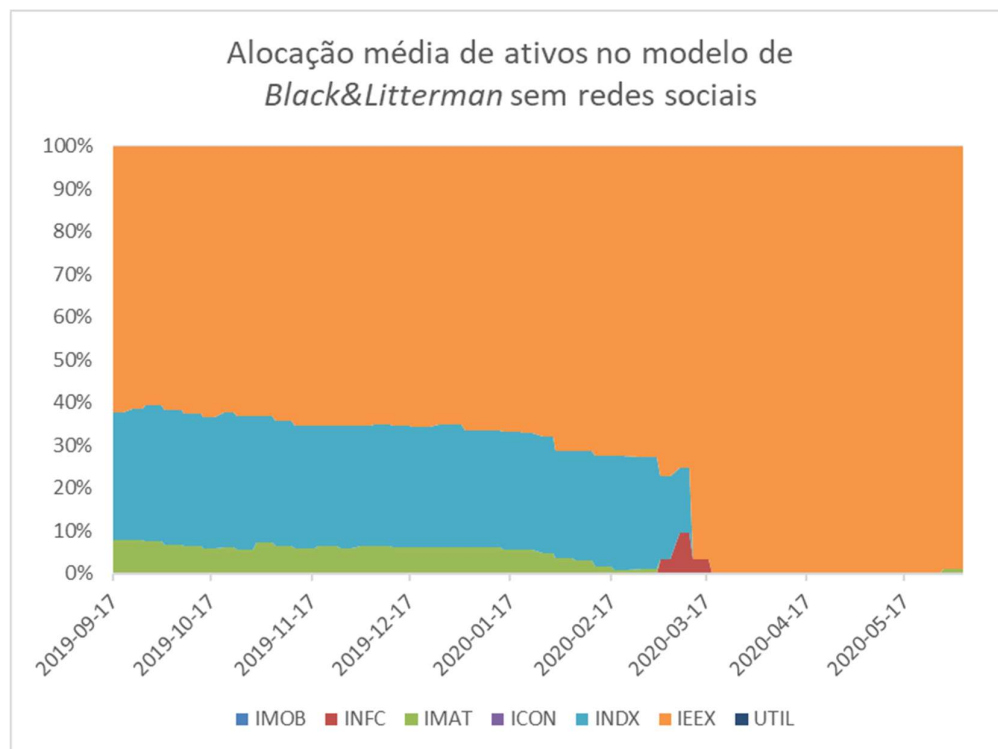


Figura 3 - Alocação média de ativos no modelo de *Black&Litterman* sem redes sociais

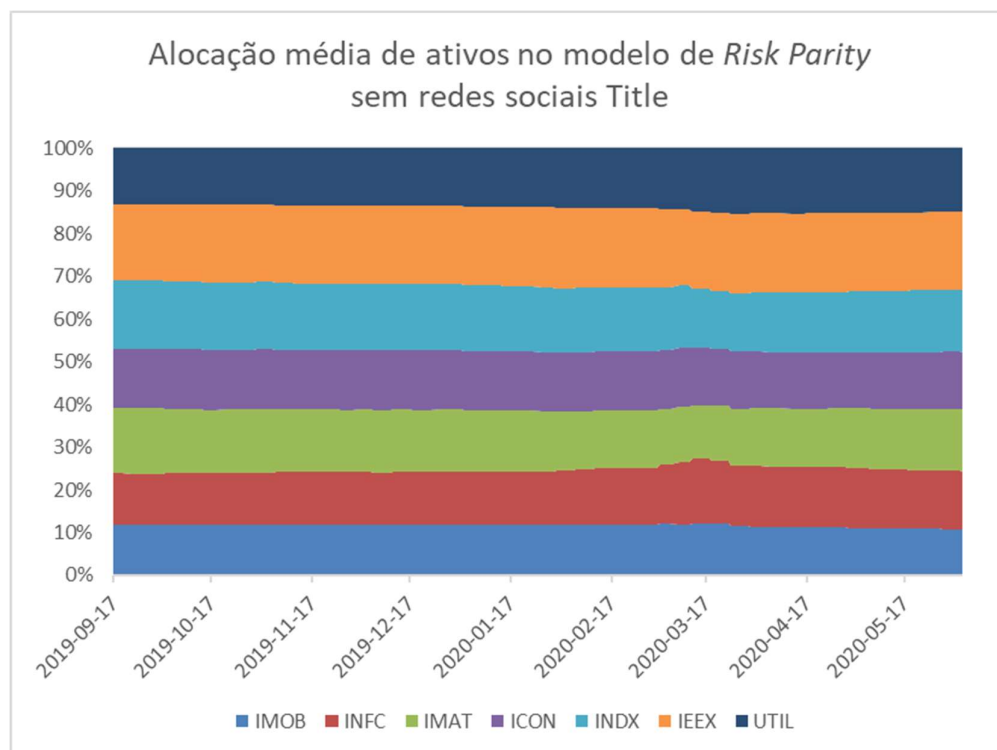


Figura 4 - Alocação média de ativos no modelo de *Risk Parity* sem redes sociais

	BL s/Social	RP s/Social	IBOV
Retorno Acumulado	9%	0%	-10%
Excesso Retorno IBOV	19,15%	10,23%	0,00%
Retorno Anualizado	12,9%	-0,1%	-14,4%
Volatilidade Diaria	3%	4%	3%
Volatilidade Anual	47,8%	56,1%	49,6%
Sharpe	0,1787	-0,0794	-0,3782

Tabela 2– Indicadores sem redes sociais

4.2. Robô de investimento com inferência de redes sociais

Vale ressaltar que o sentimento individual em relação a um determinado ativo interessa pouco para o tema de pesquisa. O interesse é em entender como os sentimentos do conjunto de ativos podem afetar a alocação. Adicionamos assim as restrições apropriadas para cada metodologia em duas intensidades: fraca ($\alpha=15\%$ e $\beta=85\%$) e forte ($\alpha=30\%$ e $\beta=70\%$).

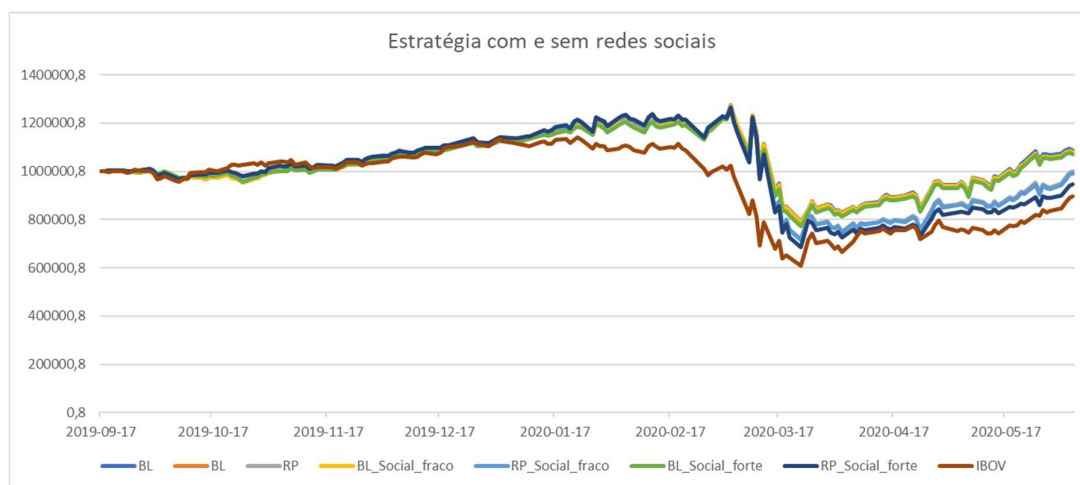


Figura 5 – Indicadores com redes sociais

A Figura 5 revela as redes sociais contribuíram negativamente para os retornos, em especial após a volatilidade de meados de março de 2020. Abaixo temos a alocação dos ativos. Vemos que houve pouca mudança em relação à alocação sem as premissas de redes sociais o que além de esperado é desejável já que as premissas não podem alterar de maneira substancial os fundamentos dos métodos de alocação.

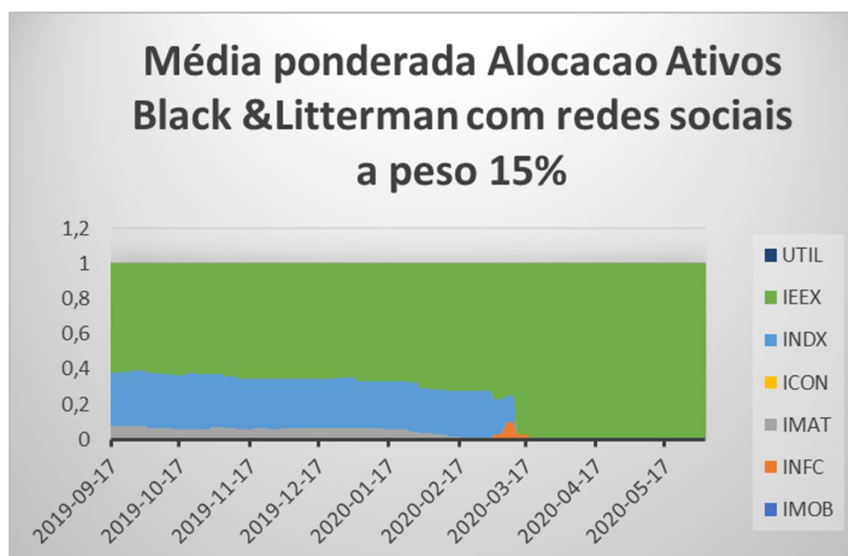


Figura 6 – Média ponderada Alocação Ativos Black&Litterman com redes sociais e Alpha de 15%

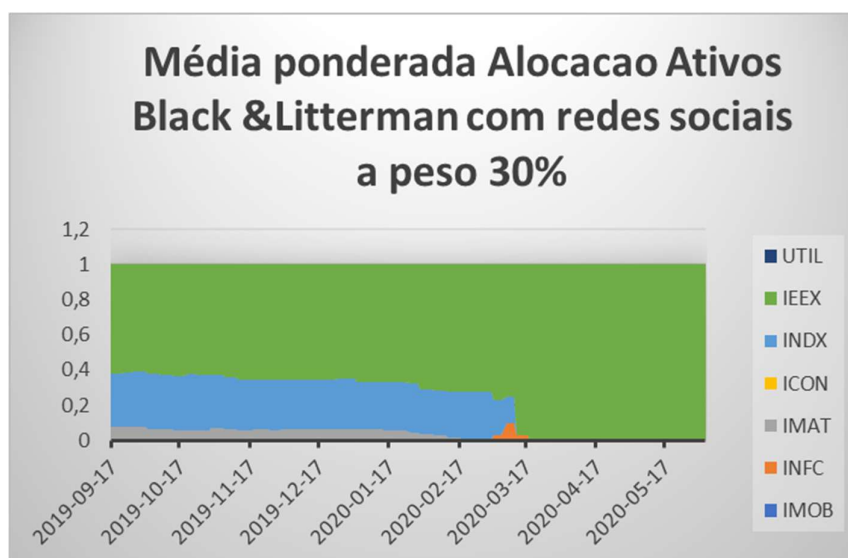


Figura 7 – Média ponderada Alocação Ativos Black&Litterman com redes sociais e Alpha de 30%



Figura 8 – Média ponderada Alocação Ativos Risk Parity com redes sociais e Alpha de 15%



Figura 9 – Média ponderada Alocação Ativos Risk Parity com redes sociais e Alpha de 30%

	IBOV	BL c/Social Fraco	RP c/Social Fraco	BL c/Social Forte	RP c/Social Forte
Retorno Acumulado	-10%	8%	-3%	7%	-5%
Excesso Retorno IBOV	0,00%	18,45%	6,90%	17,40%	5,02%
Retorno Anualizado	-14,4%	11,9%	-4,8%	10,3%	-7,5%
Volatilidade Diária	3%	3%	4%	3%	4%
Volatilidade Anual	49,6%	48,0%	56,6%	48,9%	57,6%
Sharpe	-0,3782	0,1565	-0,2	0,1216	-0,2

Tabela 3 – Indicadores com redes sociais

5. CONCLUSÕES

Podemos concluir que as técnicas de seleção de alocação de ativos apresentam resultados robustos mesmo sob cenários de alta volatilidade como tivemos em março de 2020 em razão da pandemia global.

Podemos ver retornos acima do benchmark em ambas as técnicas estudadas e índice SHARPE dentro do razoável. Nos gráficos apresentados é possível ver também que tal excedente de retorno de mantém positivo na maioria da série estudada mesmo com alguma volatilidade.

O método de paridade de risco (ou *RP*) embora tenha apresentado um retorno menor tem como grande vantagem a diversificação dos ativos, eliminando assim o risco de concentração em um setor específico. O método de *Black&Litterman* (1990) por exemplo levou a grande concentração no setor de energia, historicamente afetado por decisões regulatórias, e portanto possuiu um risco estrutural intrínseco.

Contudo as redes sociais não forneceram informações que resultassem em retornos positivos, ao contrário, diminuíram os retornos no período proposto. Tal fato corrobora grande partes dos estudos citados na bibliografia que indicam uma fraca correlação entre a interpretação dos textos de redes sociais ou notícias com os retornos dos ativos. O histórico de 1 ano, considerado curto, também pode afetar o resultado o que nos leva a inferir que uma janela maior poderia melhor o poder de estimação do modelo.

Na figura 5 podemos ver que as premissas de redes sociais afetaram muito os ativos nos períodos posteriores a alta volatilidade, em especial nos meses de fevereiro a abril, com recuperação nos meses seguintes. Tal características nos fornece alguns indícios de quais seriam as razões que levaram a esse resultado, como por exemplo as reações não racionais dos investidores amplamente estudadas em finanças comportamentais.

Finalmente podemos afirmar que o estudo fornece bons resultados em momentos de menor volatilidade de mercado, confirmando que embora tenham um potencial bastante significativo os métodos de coleta e análise dos dados precisam ser mais desenvolvidos, mantendo a premissa que a avaliação humana ainda se faz necessária, em especial nos momentos de maior volatilidade e incerteza.

6. REFERÊNCIAS

- ANBIMA. “Alocação em produtos de maior risco foram destaque em 2019 – ANBIMA”. https://www.anbima.com.br/pt_br/informar/relatorios/varejo-private-e-gestores-de-patrimonio/boletim-de-private-e-varejo/alocacao-em-produtos-de-maior-risco-foram-destaque-em-2019-8A2AB2B9701B20E901701B2D8B5F05A5.htm (28 de setembro de 2020).
- Beketov, Mikhail, Kevin Lehmann, e Manuel Wittke. 2018. “Robo Advisors: quantitative methods inside the robots”. *Journal of Asset Management* 19(6): 363–370.
- Black, Fischer, e Robert Litterman. 1990. *Asset allocation: combining investor views with market equilibrium*. Discussion paper, Goldman, Sachs & Co.
- Bourgeron, Thibault, Edmond Lezmi, e Thierry Roncalli. 2018. “Robust Asset Allocation for Robo-Advisors”.
- Cadwalladr, Carole, e Emma Graham-Harrison. 2018. “Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach”. *The guardian* 17: 22.
- Chaves, Denis, Jason Hsu, Feifei Li, e Omid Shakernia. 2012. “Efficient Algorithms for Computing RiskParity Portfolio Weights”. *The Journal of Investing* 21(3): 150–163.
- Elnagar, Ashraf, Omar Einea, e Leena Lulu. 2017. “Comparative study of sentiment classification for automated translated Latin reviews into Arabic”. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 443–448.
- “ETFGI - Independent ETFs / ETPs Research and Consultancy Firm”. *ETFGI LLP*. <https://etfgi.com/node> (15 de maio de 2020).
- “ETFs Listados | B3”. http://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/etf/renda-variavel/etfs-listados/ (15 de maio de 2020).
- “Fast Design of Risk Parity Portfolios”. <https://cran.r-project.org/web/packages/riskParityPortfolio/vignettes/RiskParityPortfolio.html#references> (22 de julho de 2020).
- Fein, Melanie L. 2015. “Robo-advisors: A closer look”.
- Feng, Yiyong, e Daniel P. Palomar. 2015. “SCRIP: Successive convex optimization methods for risk parity portfolio design”. *IEEE Transactions on Signal Processing* 63(19): 5285–5300.
- FINRA. 2016. “Report on Digital Investment Advice”. *FINRA*: 17.
- Garcia, David, e Frank Schweitzer. 2015. “Social signals and algorithmic trading of Bitcoin”. *Royal Society open science* 2(9): 150288.
- Griveau-Billion, Théophile, Jean-Charles Richard, e Thierry Roncalli. 2013. “A fast algorithm for computing high-dimensional risk parity portfolios”. *Available at SSRN 2325255*.
- Groetsch, C. W. 1984. “The theory of tikhonov regularization for fredholm equations”. *104p, Boston Pitman Publication*.
- Hoerl, Arthur E., e Robert W. Kennard. 1970. “Ridge regression: Biased estimation for nonorthogonal problems”. *Technometrics* 12(1): 55–67.

- Hutto, Clayton J., e Eric Gilbert. 2014. “Vader: A parsimonious rule-based model for sentiment analysis of social media text”. In *Eighth international AAAI conference on weblogs and social media*,.
- “Índices de Segmento | B3”. http://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-de-segmentos-e-setoriais/ (15 de maio de 2020).
- Kaya, Orçun, Jan Schildbach, Deutsche Bank AG, e Stefan Schneider. 2017. “Robo-advice—a true innovation in asset management”. *Deutsche Bank Research, August, available at https://www.dbresearch.com/PROD/DBR_INTERNET_EN-PROD/PROD000000000449010/Robo-advice_-_a_true_innovation_in_asset_managemen.pdf*.
- Ke, Zheng Tracy, Bryan T. Kelly, e Dacheng Xiu. 2019. *Predicting returns with text data*. National Bureau of Economic Research.
- Kim, Young Bin et al. 2016. “Predicting fluctuations in cryptocurrency transactions based on user comments and replies”. *PloS one* 11(8).
- Lei, Xiaojiang, Xueming Qian, e Guoshuai Zhao. 2016. “Rating prediction based on social sentiment from textual reviews”. *IEEE transactions on multimedia* 18(9): 1910–1921.
- Markowitz, Harry. 1952. “Portfolio selection”. *The journal of finance* 7(1): 77–91.
- Michaud, Richard O. 1989. “The Markowitz optimization enigma: Is ‘optimized’ optimal?” *Financial Analysts Journal* 45(1): 31–42.
- Mohammad, Saif M., Mohammad Salameh, e Svetlana Kiritchenko. 2016. “How translation alters sentiment”. *Journal of Artificial Intelligence Research* 55: 95–130.
- Nguyen, Thien Hai, Kiyooki Shirai, e Julien Velcin. 2015. “Sentiment Analysis on Social Media for Stock Movement Prediction”. *Expert Systems with Applications* 42(24): 9603–11.
- “Robo-Advisors - Worldwide | Statista Market Forecast”. *Statista*. <https://www.statista.com/outlook/337/100/robo-advisors/worldwide> (29 de maio de 2020).
- Roncalli, Thierry. 2013. *Introduction to risk parity and budgeting*. CRC Press.
- . 2014. “Introducing expected returns into risk parity portfolios: A new framework for asset allocation”. *Available at SSRN 2321309*.
- Sharpe, William F. 1964. “Capital asset prices: A theory of market equilibrium under conditions of risk”. *The journal of finance* 19(3): 425–442.
- . 1992. “Asset allocation: Management style and performance measurement”. *Journal of portfolio Management* 18(2): 7–19.
- Tibshirani, Robert. 1996. “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1): 267–288.