B3 Stock Price Prediction Using LSTM Neural Networks and Sentiment Analysis

Gabriel M. Vargas, Leonardo J. Silvestre, Luís O. Rigo Jr. and Helder R. O. Rocha

***Abstract*—This article presents an approach to predict stock prices which incorporate sentiment analysis from Twitter posts as an input to an Long Short Term Memory (LSTM) Neural Network to help in the decision process. The sentiment analysis measures subjectivity and polarity as well as the number of tweets about the company to capture the market mood, which influences the stock prices, were evaluated. The main company used to evaluate our method is Vale (VALE3). The sentiment analysis helps to reach an Root Mean Squared Error (RMSE) of 0.021. We also validate our method with JHSF (JHSF3) and Usiminas (USIM3), obtaining RMSE of 0.012 and 0.016, respectively.**

***Index Terms*—Financial Market, LSTM, Recurring Neural Networks, Sentiment Analysis, Stock Exchange.**

I. Introdução

A aplicação de recursos financeiros é uma tarefa complexa para os investidores, devido à quantidade de informações oriundas de diferentes contextos [1]. Toda empresa ou investidor individual, ao aplicar seus recursos, considera pelo menos os aspectos do retorno e do risco. Com um nível adequado de informações e conhecimento do mercado financeiro é possível, para certo nível de retorno, reduzir a exposição ao risco [2].

No entanto, tentar prever o comportamento do preço de uma ação na bolsa de valores é uma tarefa bastante complexa, pois seu preço tem um comportamento muito dinâmico. Várias variáveis, que vão desde a emoção coletiva a notícias de alto nível, passando ainda por variáveis desconhecidas, influenciam nos valores de uma ação. Essa complexidade leva a uma volatilidade que pode ser perigosa e trazer grandes perdas para investidores, principalmente os inexperientes. Mesmo assim, o dinamismo das ações atrai a atenção dos investidores, pois os investimentos podem proporcionar bons lucros quando feitos no momento certo e da maneira correta [3].

Nesse sentido, todas as possibilidades de minimizar a complexidade e o risco dos investimentos são bem-vindas. Com isso, no sentido de tentar antecipar as mudanças do mercado fazendo uso de dados coletados previamente, é possível aplicar técnicas de Inteligência Artificial como as Redes Neurais Artificiais (RNAs) e a análise de sentimentos como ferramentas de apoio à tomada de decisão. As RNAs, que são modelos matemáticos cuja arquitetura e funcionamento baseiam-se na forma que o cérebro humano trabalha, possuem a capacidade de reconhecer padrões, identificar regularidades, lidar com dados ruidosos, incompletos ou imprecisos e de prever sistemas não lineares, o que torna a sua aplicação interessante no mercado financeiro [4] e [5].

Baseado na premissa de que as notícias e acontecimentos podem influenciar de forma substancial no preço das ações, diversas técnicas que visam rotular os textos para que seja possível qualificá-los como bons e ruins, relevantes ou não, têm ganhado espaço. Desta forma, o uso da análise de sentimentos tem se tornado um importante tópico na Web, especialmente em redes sociais, com o desenvolvimento de aplicações para monitoramento de produtos e marcas, assim como na análise da repercussão de eventos [6]. Opiniões nas redes sociais, se devidamente recolhidas e analisadas, permitem não só compreender e explicar diversos fenômenos sociais complexos, mas também prevê-los [7]. Vários métodos e técnicas vêm sendo propostos na literatura e sua aplicação na predição de ações no mercado financeiro é de extrema valia [8]. Diversas pesquisas vêm sendo desenvolvidas para a predição de ações [9]. No entanto, muito ainda pode ser estudado para que as técnicas apresentadas se tornem ferramentas práticas e baratas no apoio à tomada de decisão no mercado financeiro.

A. Trabalhos Relacionados

Nos últimos anos, vários pesquisadores tentaram fazer uso de alguma técnica de inteligência artificial para prever os valores de ações do mercado financeiro. Mais recentemente, alguns trabalhos enfatizam o uso de análise de sentimentos para auxiliar nessa previsão [10], [11], [12], [13], [14]. Em [6] foi investigado se as medidas de estados de humor coletivos derivados dos *tweets* em grande escala estão correlacionados ao valor do Dow Jones ao longo do tempo. Os resultados indicaram que a precisão das previsões do Dow Jones poderia ser significativamente melhorada pela inclusão das informações específicas do humor público. Liu [5] propõe um modelo para analisar o sentimento de fórum de ações online e usar as informações para prever a volatilidade das ações. Nesse trabalho, os autores rotularam os sentimentos de publicações financeiras online e combinaram com os dados do mercado para a previsão da volatilidade das ações usando as Redes Neurais Recorrentes (*Recurrent Neural Networks* - RNNs). O modelo com indicadores sentimentais foi significativamente melhor quando comparado ao modelo de RNN sem tais indicadores.

Borovkova e Tsiamas [4] apresentaram um conjunto de redes neurais de memória de longo-curto prazo (*Long Short*

Gabriel M. Vargas, Universidade Federal do Espírito Santo (UFES), São Mateus, Espírito Santo, Brasil, gab.mv@hotmail.com

Leonardo J. Silvestre, Universidade Federal do Espírito Santo (UFES), São Mateus, Espírito Santo, Brasil, leonardo.silvestre@ufes,br

Luís O. Rigo Jr., Universidade Federal do Espírito Santo (UFES), São Mateus, Espírito Santo, Brasil, luis.rigo@ufes.br

Helder R. O. Rocha, Universidade Federal do Espírito Santo (UFES), Vitória, Espírito Santo, Brasil, helder.rocha@ufes.br

Term Memory - LSTM) para previsões de ações, usando uma grande variedade de indicadores de análise técnica como entradas de rede. Avaliam o modelo proposto sobre várias ações de grande capitalização dos EUA, obtendo um melhor desempenho que o *benchmark*. Em [15], são apresentados três diferentes modelos de rede neural, que são comparados com dados brutos, dados processados com média móvel simples e dados filtrados com transformação *wavelet* discreta. Aplicar a transformação *wavelet* nos dados possibilitou à LSTM obter melhores resultados.

Ko [10] utilizou artigos de notícias e ferramenta de processamento de linguagem natural *Bidirectional Encoder Representations from Transformers* (BERT) para reconhecer os sentimentos dos textos. Nesse trabalho, a rede neural LSTM foi usada para analisar dados de séries temporais, sendo aplicada para prever o preço das ações com informações de transações históricas e sentimentos de texto. De acordo com os resultados experimentais usando os modelos propostos, a raiz do erro médio quadrático (RMSE) tem 12,05% de melhoria na precisão. Em [16] também usou-se análise de sentimento em notícias relacionadas com o mercado e ações e, em seguida, combinou os rótulos de sentimento das notícias com os preços das ações e indicadores técnicos para treinar uma rede neural LSTM em que a saída da rede é usada na previsão do movimento do preço de fechamento. Os experimentos obtiveram uma pontuação F1 de 0,65 quando os preços das ações são combinados com rótulos de sentimento.

Já em [17] é apresentada uma revisão da literatura sobre o uso de métodos de Aprendizado de Máquina (AM) para a previsão do mercado de ações. Os trabalhos são classificados com base em diferentes técnicas de predição e agrupamento. As técnicas comumente usadas para obter uma previsão eficaz do mercado de ações são a RNA e a abordagem *fuzzy*. Mesmo com muitos esforços de pesquisa, a técnica atual de previsão do mercado de ações ainda tem muitos limites. Pode-se concluir que a previsão do mercado de ações é uma tarefa muito complexa, e diferentes fatores devem ser considerados para prever o futuro do mercado de forma mais precisa e eficiente. Em [11], foram usados algoritmos em mídia social e dados de notícias financeiras para descobrir o impacto desses dados na precisão da previsão do mercado de ações. Compararam-se os resultados de diferentes algoritmos para encontrar um previsor consistente. Os resultados mostram que as melhores previsões são obtidas por meio das mídias sociais e notícias financeiras, respectivamente com precisões de 80,53% e 75,16%. A técnica de floresta aleatória foi considerada consistente e a maior precisão encontrada com ela foi de 83,22%.

O trabalho em [12] utiliza a análise de sentimentos do investidor e o índice de ações para prever a variação do preço das ações. O processo de extração de sentimento sobre os posts textuais gera 6 variáveis que compõem índices de sentimento dos investidores. Os autores afirmam que o método proposto pode melhorar a precisão dos métodos de AM na previsão de tendência do índice de ações. Afirmam, também, que o SVM obteve melhor resultado.

Enquanto que em [13] é estudada a influência das notícias publicadas nos valores das ações das empresas, através da análise de sentimento presente nessas notícias. Os autores realizaram experimentos usando o classificador *Naive Bayes*, MLP e a abordagem lexical. Os resultados mostram que os modelos de aprendizado de máquina superam a abordagem lexical.

Em [18], é aplicado um modelo LSTM para classificar sentimentos de *tweets* em positivo e negativo com relação a preços de ações. O trabalho, portanto, é focado em classificação, apresentando métricas relacionadas com essa tarefa.

Já em [14] analisou-se o impacto das variáveis de sentimento no mercado de ações usando dados que combinam mídias sociais, notícias, artigos e dados de mecanismos de pesquisa. Foi usada a técnica de classificação auto-regressiva heterogênea para predizer a volatilidade do mercado de ações. O estudo mostrou que as variáveis de sentimento são capazes de melhorar significativamente as previsões das ações.

B. Contribuições do Trabalho

Neste trabalho, apresenta-se o desenvolvimento de uma metodologia capaz de auxiliar na tomada de decisão de compra e venda de ações da bolsa de valores baseando-se em fusão de informações colhidas do *Twitter* (https://twitter.com/) em um espaço de tempo pré-determinado, valores históricos das ações, índices e câmbio de moedas. Foi avaliada também a relevância de cada uma das informações na previsão que auxilia na tomada de decisão. Para isso, foi utilizada uma abordagem de rede neural recorrente, a LSTM, mais apropriada para séries temporais. Para tratamento dos dados do *Twitter*, utilizou-se a técnica de análise de sentimento, permitindo agregar estes dados ao conjunto de entrada da rede neural, com o objetivo de prever o preço médio da ação no pregão para o dia seguinte. Foi desenvolvida uma metodologia para extração de características dos valores resultantes da análise de sentimento, de forma a permitir inserir os valores de sentimento como entrada da rede, em conjunto com os dados históricos da B3. Nesse sentido, foram testadas três métricas para quantificar o sentimento. Complementarmente, foram realizados experimentos comparando o resultado da LSTM com uma rede perceptron de múltiplas camadas (*Multi-layer perceptron* - MLP), assim como a validação do modelo obtido para diferentes ações da bolsa.

O presente artigo está organizado da seguinte maneira. Na Seção II, são apresentados os principais conceitos sobre mercado financeiro, sobre a técnica de rede neural utilizada e também sobre análise de sentimento. Em sequência, na Seção III, será descrita a metodologia experimental utilizada neste trabalho. Na Seção IV serão apresentados os resultados obtidos e as discussões sobre eles. Por fim, a Seção V apresenta as principais conclusões baseadas nos resultados obtidos e possíveis trabalhos futuros.

II. Referencial Teórico

A. Mercado Financeiro

O mercado financeiro é o local onde podem ser negociados, comprados ou vendidos, bens como valores mobiliários, mercadorias e câmbio. Para realizar tais negociações, várias instituições financeiras estão envolvidas, facilitando o encontro entre as partes, regulando e fiscalizando as transações. Hoje,

no Brasil, há uma empresa responsável pela bolsa de valores, que administra e registra as negociações de diferentes tipos de ativos e fornece liquidez para o mercado financeiro, conhecida como B3 (Brasil, Bolsa, Balcão) [19].

Existem vários ativos (tudo aquilo que é negociado no mercado financeiro) negociados na bolsa, sendo fundos imobiliários, fundos de renda fixa, fundos de renda variável e também as ações, que nada mais são que fragmentos de empresas cujo capital é aberto. Estas ações podem ser negociadas na bolsa em lotes de 100 ou no mercado fracionário, onde podem ser negociadas uma por vez, e podem ser preferenciais (PN) ou ordinárias (ON). Todas essas negociações são feitas entre o horário de abertura e fechamento do pregão eletrônico, que é basicamente o substituto do pregão tradicional, conhecido como viva voz. Os preços de um ativo na bolsa podem variar por diversos fatores, sendo eles conhecidos ou não, tornando a tarefa de prever seus valores muito complexa. A forma como os valores oscilam dia após dia pode ser observada como uma série temporal [9].

O investimento em mercado de ações segue duas linhas de pensamento. A primeira, denominada de análise fundamentalista de mercado, é a avaliação de empresas a partir dos seus indicadores de endividamento, retorno e liquidez para encontrar possíveis erros de precificação. A segunda, objetivo deste trabalho, é a análise técnica, que é definida como uma metodologia de previsão de tendências baseada apenas em dados históricos, principalmente de preço e volume [19].

B. Redes Neurais Recorrentes para Previsão de Tendências

As RNAs possuem capacidade de lidar com não-linearidades, reconhecer padrões, identificar regularidades, lidar com dados ruidosos, e apresentam resultados robustos para aplicações complexas nos mais diversos domínios. Dada a complexidade e a característica de comportamento temporal do mercado de ações, as RNAs, em especial as redes LSTM, vêm sendo utilizadas com o objetivo de prever o comportamento dos preços das ações.

LSTM é um modelo recorrente e profundo de redes neurais amplamente usado e capaz de alcançar alguns dos melhores resultados quando comparado com outros métodos aplicados à previsão de séries temporais [20], com destaque para o campo de Processamento de Linguagem Natural [21]. Esse modelo surgiu para resolver o problema de desaparecimento de gradiente, presente em redes neurais recorrentes padrão, quando lidam com a dependência temporal sob longas sequências de dados.

Assim, além dos componentes padrão que simulam a recorrência em uma RNN e que são suscetíveis ao desaparecimento do gradiente, as LSTMs possuem componentes especiais, chamados "portões" (*gates*), que permitem esquecer uma informação que não é mais relevante ou adicionar uma nova informação na memória [22].

A Figura 1 apresenta o modelo esquemático de uma LSTM contendo um único módulo, em que a dependência temporal é ilustrada pela repetição sequencial deste módulo ao longo do tempo(da esquerda para direita). O módulo central da figura apresenta um detalhamento dos componentes e do fluxo de informações em uma rede LSTM. Cada linha preta representa o sentido do fluxo de um conjunto de informações. As circunferências em rosa representam operações pontuais, como uma adição de vetor, enquanto os retângulos amarelos são conjuntos de neurônios com um tipo específico de função de ativação. A fusão das linhas significa concatenação, enquanto uma bifurcação de linha implica que seu conteúdo foi copiado e suas cópias estão indo para locais diferentes.

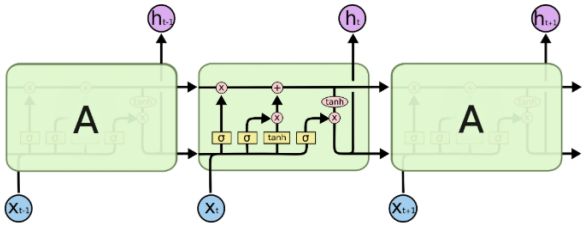

Fig. 1. Diagrama esquemático de uma LSTM [23].

Um módulo LSTM recebe três conjuntos de informações para realizar o processamento: vetor de entradas X; última resposta da rede; memória armazenada na rede. O vetor de entrada e a última resposta da rede são utilizados para definir o comportamento dos portões. O **portão de esquecimento** (*forget gate*) é responsável por decidir quais informações da memória devem ser descartadas ou mantidas. O **portão de entrada** (*imput gate*) é responsável por atualizar o estado da memória adicionando novas informações. O último portão, **portão de saída** (*output gate*), decide qual deve ser a saída h da rede LSTM. Este fluxo de informações é repetido sequencialmente ao longo do tempo. Além disso, os módulos LSTM podem ser empilhados para formar uma rede mais profunda.

C. Análise de Sentimento

Análise de sentimento é um ramo de mineração de textos para classificar textos pelo sentimento ou opinião contidos no mesmo. O termo é mais utilizado para significar o tratamento computacional da opinião, sentimento e subjetividade em textos [24]. No âmbito textual, uma notícia pode ser classificada em positiva ou negativa sem ser de opinião, como por exemplo, na frase "esta empresa está crescendo de uma forma muito saudável", o sentimento para essa notícia seria classificado como positivo pois, claramente, trata-se de uma notícia boa. A classificação do sentimento dá-se em valores numéricos, normalmente um número entre -1 (muito negativo) e 1 (muito positivo). Isso dá maior eficiência para a análise. Há ainda a possibilidade de se classificar em classes como Positivo, Neutro e Negativo, além de variações dessas classes.

A Figura 2 apresenta o resultado da análise da frase "O dia está lindo!" na ferramenta de análise de sentimento, que faz parte da API de linguagem natural do Google (https://cloud.google.com/natural-language?hl=pt-br). Ela retorna o nível de sentimento a partir da entrada fornecida. Pode-se observar que o valor resultante (*Score*) foi 0,9, indicando um sentimento bastante positivo para a frase analisada.

A análise de sentimentos tem sido usada em diferentes segmentos, por exemplo:

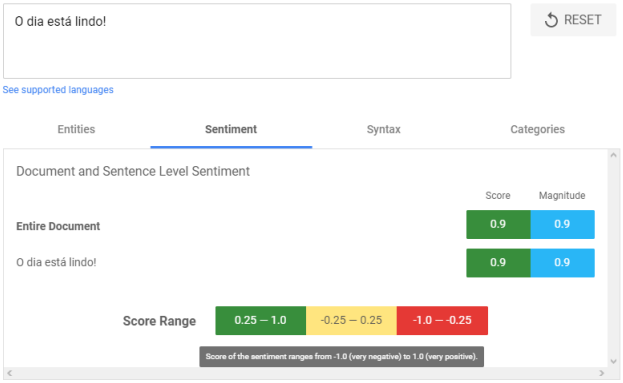

Fig. 2. Análise do sentimento para a frase “O dia está lindo!”.

- Política: para medir a popularidade de um determinado candidato a cargo público eletivo;
- Indústria: para avaliar a aceitação pelos consumidores de um determinado produto;
- Bolsa de Valores: para medir o clima coletivo em determinado ativo negociado na bolsa de valores;
- Pessoas comuns: para pesquisar opiniões sobre assuntos públicos e pessoas.

A partir de 2010, a evolução das redes sociais, em especial o *Twitter*, tornou possível prever uma variedade de temas através das mensagens trocadas entre utilizadores. Entre esses temas, estão a classificação de produtos e o resultado de eleições, mostrando que o *Twitter* pode ser uma boa fonte para extração de indicadores e previsões [25]. Para exemplificar, a Figura 3 apresenta o valor da ação VALE3 entre os anos de 2013 e 2018, assim como a quantidade de *tweets* no mesmo período. Pode-se observar que, entre o fim de 2015 e o início de 2016, houve uma queda grande no valor da ação e um aumento significativo na quantidade de *tweets*, refletindo o desastre da barragem de Mariana-MG. Pode-se observar, ainda, que no ano de 2016 não há *tweets* pois, apesar de terem sido feitas requisições, a API utilizada não retornou resultados.

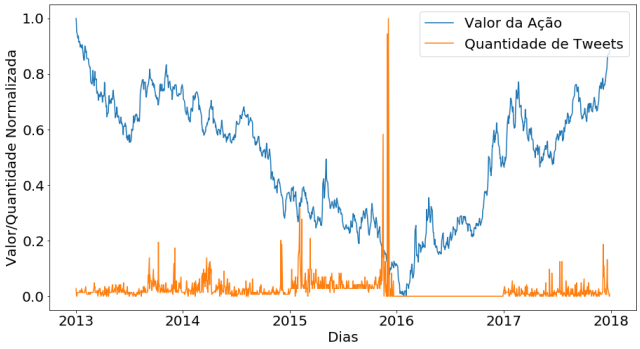

Fig. 3. VALE3 - Valor da ação × Quantidade de *tweets* normalizados.

III. Metodologia

A. Geração do Dataset das Variáveis Econômicas

Os dados utilizados neste trabalho são de uso público e podem ser obtidos nos sites da B3 (http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/historico/mercado-a-vista/cotacoes-historicas/), YAHII (http://www.yahii.com.br) e do *Twitter*. Da B3 foram obtidas as informações das cotações históricas relativas à negociação de todos os ativos para cada ano, que futuramente seriam filtrados para a empresa Vale, JHSF e Usiminas, referente as ações ordinárias (ON). O período escolhido para a análise foi de 5 anos, de 2013 a 2017, que no caso da empresa Vale inclui o desastre da barragem de Mariana-MG. Os dados históricos obtidos do site da B3 possuem um conjunto grande de atributos, mas para este trabalho só serão utilizados 7 deles, que são:

- PREABE: Preço de abertura do papel no pregão;
- PREMAX: Preço máximo do papel no pregão;
- PREMIN: Preço mínimo do papel no pregão;
- PREMED: Preço médio do papel no pregão;
- PREULT: Preço do último negócio do papel no pregão;
- PREOFC: Preço da melhor oferta de compra do papel;
- PREOFV: Preço da melhor oferta de venda do papel.

Os dados do YAHII são referentes às outras variáveis que farão parte do conjunto de entrada da rede neural, como variações cambiais do Dólar, taxa de juros SELIC, índice BOVESPA (Ibovespa) e também do Dow Jones. Com todos os dados em mãos, os dados históricos da ação da Vale foram acrescidos dos dados históricos das variáveis de mercado.

A Figura 4 apresenta o mapa de calor, análise que indica a correção entre os valores das variáveis presentes na série temporal. Pode-se observar que existem algumas correlações fortes entre algumas variáveis. Por exemplo, o valor de Down Jones tem forte correlação com Ibovespa, assim como abertura tem correlação muito forte com fechamento. Outra correlação importante, em especial para este trabalho, é a de Ibovespa com Vale. O próximo passo foi fazer a análise de sentimento.

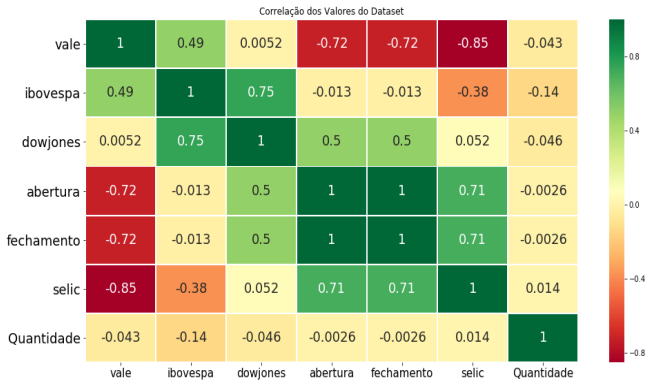

Fig. 4. Mapa de calor da correlação dos dados.

B. Geração do Dataset para Análise de Sentimento

1) Obtenção dos Tweets*:* do *Twitter*, foram obtidas postagens (*tweets*) com informações relativas à economia e a notícias que têm relação com as empresas selecionadas. Os *tweets* foram coletados usando uma API fornecida pelo próprio

Twitter, por meio de uma conta de desenvolvedor concedida para fins de pesquisa. Com esta conta, foi possível obter acesso a todos os *tweets* desde a criação da empresa, limitado a 50 requisições mensais, sendo que em cada requisição era possível coletar 100 *tweets*. Foram extraídos 4809 *tweets* dos perfis da Vale (@valenobrasil), do jornal Estadao (@Estadao), do portal de notícias InfoMoney (@infomoney) e do portal de notícias G1 (@g1). Todos os dados foram extraídos utilizando a API do *Twitter* usando as *hashtags* (#) "Vale" e "Economia".

2) Tradução dos tweets para o Inglês e cálculo dos valores de sentimento: foi utilizada a API *googletrans* para traduzir os *tweets* para o inglês, que é o idioma utilizado pela *textblob*, biblioteca que trabalha com processamento de linguagem natural (PLN) por meio de outra biblioteca de PLN, a *Natural Language Toolkit*, que permite acesso fácil a muitos recursos léxicos e permite trabalhar com categorização, classificação e também suporta análises mais complexas em dados de texto. A sentença traduzida é, então, convertida para o formato tipo *textblob*.

3) Cálculo dos valores de sentimento: com os *tweets* traduzidos, passou-se para a etapa de análise de sentimento, através de um algoritmo capaz de calcular a polaridade, que significa o quão positivo ou negativo o *tweet* pode ser, variando de -1 até 1, e a subjetividade, que varia de 0 a 1. Após o cálculo, o sentimento e a subjetividade de cada sentença ficam acessíveis por meio de métodos.

4) Extração de características dos sentimentos: com o *dataset* dos *tweets* completamente analisado e com os sentimentos e subjetividade gravados, esse resultado é acrescido aos dados da ação. Como a quantidade de *tweets* por dia pode variar de 0 a N, e a informação da ação neste exemplo é obtida por dia, ou seja, possui somente uma entrada, foram testadas três formas de transformar os dados para que ficassem em um formato compatível com os da ação. Estas são as três possibilidades:

1) Na primeira, para evitar divisão por zero, soma-se um aos valores da subjetividade S, que passam a variar agora entre 1 e 2. Então, a polaridade P é dividida pela subjetividade, formando um novo campo de sentimento $NSent_1$. No fim, somam-se os n valores desse campo na data desejada, onde n é a quantidade de *tweets* na data. A Equação 1 apresenta esse cálculo.

$$NSent_1 = \sum_{i}^{n} P_i/(S_i + 1) \tag{1}$$

A ideia por trás desta metodologia é que, quanto menos subjetivo é o comentário, ou seja, quanto mais próximo de zero, maior relevância ele deve ter, tanto para polaridade positiva quanto para negativa. Caso o comentário seja 100% objetivo, ou seja, com subjetividade igual a um, seu valor de sentimento deve ser mantido, e caso a subjetividade esteja próxima de 2, seu valor será reduzido, diminuindo a influência do mesmo no somatório;

2) Na segunda, os valores calculados para $NSent_1$ (Equação 1) são categorizados em 'Bom', 'Ruim' e 'Neutro', conforme Equação 2:

$$NSent_2 = \begin{cases} -1 \leq NSent_1 < -0.4 & = Ruim \\ -0.4 \leq NSent_1 < 0.4 & = Neutro \\ 0.4 \leq NSent_1 \leq 1 & = Bom. \end{cases} \tag{2}$$

3) Na terceira abordagem, é feita a soma dos campos de polaridade e subjetividade de forma independente, resultando em dois novos campos: soma da polaridade na data e soma da subjetividade na data.

As três possibilidades foram testadas na rede de melhor desempenho e a que apresentou o melhor resultado no teste empírico foi adotada. A Figura 5 apresenta o fluxograma da metodologia utilizada na análise de sentimento.

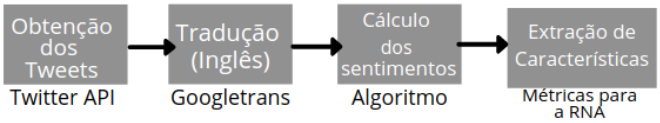

Fig. 5. Fluxograma da metodologia utilizada na análise de sentimento.

C. Tratamento dos Dados

Com os dados todos unificados em uma mesma estrutura, o primeiro passo foi fazer uma cópia dos valores do preço médio (PREMED), pois esse será o valor a ser previsto, porém com uma adaptação: cada amostra do *dataset* foi formada por todos os valores do dia atual (utilizado como dado de entrada da rede) + o preço médio (PREMED) do dia seguinte (valor a ser previsto). Com isso, o último dia do *dataset* de exemplos foi eliminado e apenas o preço médio deste dia foi usado como valor previsto do dia anterior.

Após esse procedimento, o *dataset* foi normalizado no intervalo de [0,1] e dividido sequencialmente na proporção de 80% para treinamento e 20% para teste. Dos dados de treinamento, 20% foram separados para validação.

D. Criação da Rede Neural

Dado que o problema envolve séries temporais, sendo importante a memória de fatos anteriores para a previsão, faz-se necessário o uso de uma rede recorrente, tendo sido utilizada uma LSTM. A rede neural foi criada utilizando a biblioteca Keras, com o ajuste da arquitetura e dos hiperparâmetros da rede neural feita de forma experimental (empírica). A Figura 6 apresenta o fluxograma da metodologia utilizada na LSTM.

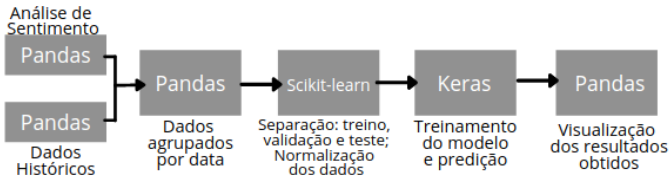

Fig. 6. Fluxograma da metodologia utilizada na LSTM.

A Tabela I apresenta a arquitetura final, escolhida após experimentos preliminares de calibração do modelo. Nessa, após cada camada LSTM, há uma camada de *dropout* (valor de

0,25). FC refere-se à camada totalmente conectada (*Fully Connected*). Os outros hiperparâmetros são: tangente hiperbólica como função de ativação; MSE como função de perda; e adam como otimizador.

TABELA I
ARQUITETURA ESCOLHIDA PARA A LSTM.

Camada	Tipo	#Neur.	Camada	Tipo	#Neur.
#1	Entrada	8			
#2	LSTM	100	#5	LSTM	50
#3	LSTM	100	#6	FC	50
#4	LSTM	48	#7	Saída	1

Após experimentos preliminares, definiu-se o número de épocas e tamanho do *batch* como 50 e 64, respectivamente, para as ações da Vale.

Finalizada a definição do número de épocas e do tamanho do *batch*, os experimentos para calibração dos demais hiperparâmetros foram realizados 10 vezes para cada configuração de *dataset*. A melhor rede de cada experimento foi escolhida usando *checkpoint*, isto é, a rede com o melhor valor de validação foi salva durante o treinamento. Por fim, a melhor rede para cada configuração de *dataset* foi usada para teste.

E. Setup Experimental

Todo o processo de desenvolvimento foi feito no ambiente Google Colaboratory, utilizando linguagem Python. A plataforma disponibiliza processador Intel Xeon com 2GHz, 12GB de RAM e GPU que varia entre os modelos Tesla K80, Tesla P100 e Tesla T4, de acordo com a disponibilidade momentânea.

IV. RESULTADOS EXPERIMENTAIS

A. Métricas de Análise de Sentimento

A Figura 7 apresenta os sentimentos relacionados à Usiminas entre os anos de 2013 e 2018, classificados entre Ruim, Neutro e Bom. Os períodos sem *tweets* são requisições à API que não retornaram dados. Observa-se que, no primeiro semestre de 2016, ocorreu uma grande quantidade de *tweets*, inicialmente "neutros" e posteriormente "ruins". Essa mudança para "ruins" tem relação com notícias de prejuízos após divulgação de balanço trimestral da empresa. Outro pico de *tweets* aconteceu no segundo semestre de 2017. Nesse caso, prevalecem sentimentos classificados como "bons" e "neutros", o que tem relação com notícias de lucros no período. Ou seja, os valores extraídos para os sentimentos apresentam forte relação com acontecimentos que refletem nos valores das ações.

B. Ação Ordinária da Vale (VALE3)

1) LSTM: Objetivou-se nesta subsecção avaliar quais dados de mercado teriam maior influência no resultado da rede para a ação da Vale, considerando a rede LSTM. Inicialmente, verificou-se o resultado utilizando apenas os dados históricos, sem utilização de sentimentos e outros dados. O RMSE obtido para esses dados foi de 0,025. Posteriormente, a rede foi

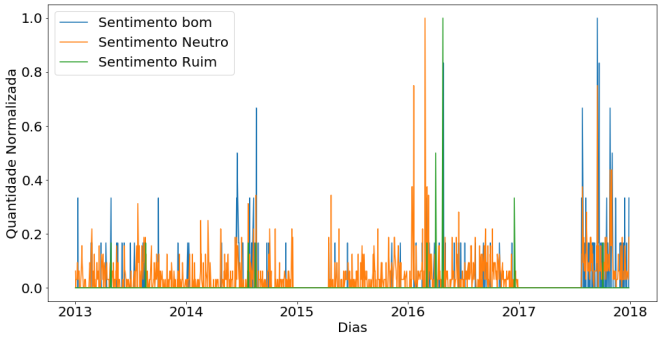

Fig. 7. USIM3 - Série temporal da quantidade de *tweets* classificados em "bom", "neutro" e "ruim", com valores normalizados.

treinada com os dados da Vale em conjunto com o Ibovespa, resultando em um RMSE de 0,021. Após, foram testadas as três variações de sentimentos apresentadas na Subseção III-B, obtendo RMSEs de 0,0218, 0,0213 e 0,0216 para as formas 1, 2 e 3, respectivamente. Foram testados, ainda, os dados da Vale em conjunto com o Ibovespa e com a Selic, obtendo um RMSE de 0,030; em conjunto com Ibovespa e com a cotação do dólar, com RMSE de 0,063 e em conjunto com Ibovespa e Dow Jones, com RMSE de 0,024. As Figura 8 e 9 apresentam, respectivamente, a evolução do aprendizado (função de perda sobre treinamento e validação) e os valores previstos para as ações da VALE3 em conjunto com a forma 2 de análise de sentimento, que gerou o melhor valor de RMSE.

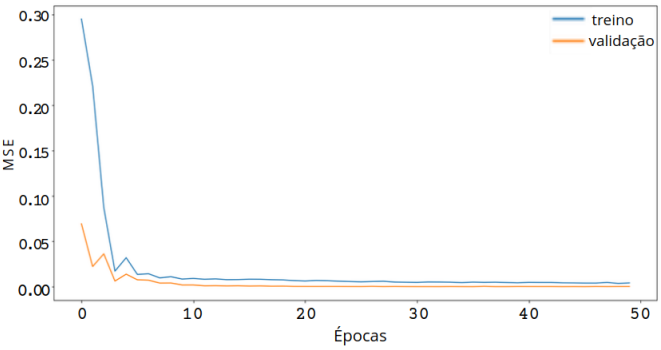

Fig. 8. Evolução da função de perda (MSE) da rede LSTM durante o aprendizado.

É interessante observar (Figura 9) que a rede consegue acompanhar muito bem as tendências da ação, apresentando um distanciamento apenas no mínimo global, cujo comportamento atípico foi causado devido ao desastre da barragem de Mariana-MG.

2) MLP: foi utilizada uma rede MLP, com 100 neurônios na camada escondida e *relu* como função de ativação, com o objetivo de analisar o desempenho de uma arquitetura não-recorrente, comparando com o modelo recorrente utilizado no presente trabalho. Para tal, a MLP recebeu os dados do Ibovespa e a da análise de sentimento da VALE na forma 2 como entrada, obtendo RMSE de 0,041.

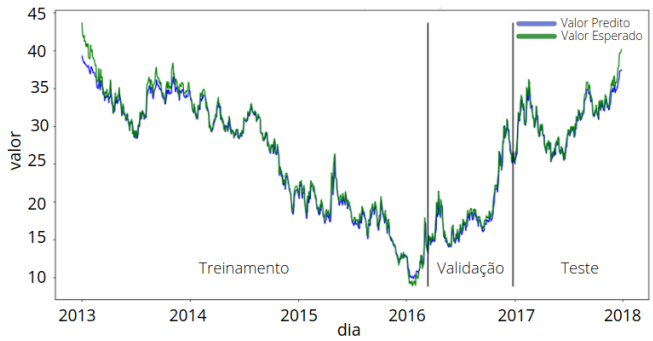

Fig. 9. Previsões da LSTM para VALE3 utilizando análise de sentimento na forma 2.

C. Ações Ordinárias da JHSF e Usiminas

Por fim, a rede LSTM foi avaliada em outras duas ações: JHSF (JHSF3) e Usiminas (USIM3). Para a JHSF, a única alteração no treinamento da rede foi o número de épocas, que foi alterado para 70. O melhor conjunto de informações para as ações da JHSF foi de 0.012 de RMSE, obtido apenas com os dados históricos da ação, tendo em vista que não havia *tweets* disponíveis para realizar a análise de sentimento, já que o limite disponibilizado pela API havia sido excedido. Já para a Usiminas, foi realizado um ajuste no tamanho do *batch* para 110, obtendo o melhor resultado com os dados históricos da ação em conjunto com os dados do Ibovespa e os sentimentos na forma 2, obtendo RMSE de 0,016. É interessante observar que foi possível obter resultados melhores para ativos, considerando uma mesma arquitetura de LSTM. A Figura 10 apresenta o gráfico com os valores reais da ação USIM3 e as previsões feitas pela LSTM.

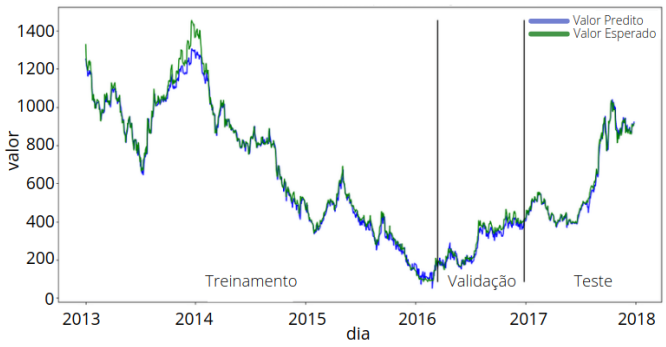

Fig. 10. Previsões da LSTM para USIM3 utilizando análise de sentimento na forma 2.

A Tabela II apresenta o resumo dos melhores resultados de teste encontrados para cada *dataset* testado, considerando a mesma arquitetura de rede neural LSTM. O melhor resultado para cada ação está destacado, em negrito. Além da tradicional métrica RMSE, é apresentado ainda o valor da métrica *U de Theil*. Tal métrica compara os valores obtidos com o método ingênuo, que consiste em apenas utilizar o valor medido no dia d como previsão para o dia $d+1$. Valor de *U de Theil* menores que um significam que o método é melhor que o ingênuo [26]. Como pode ser observado para os resultados da VALE3, a forma 2 de análise de sentimento obteve, além do menor RMSE, um valor de *U de Theil* abaixo de zero, mostrando que a previsão obtida é melhor do que o método ingênuo e melhor que o resultado dos demais experimentos. Isso se repete para as previsões obtidas para JHSF e Usiminas.

TABELA II
RESUMO DE RESULTADOS DA LSTM SOBRE OS DIVERSOS DATASETS.

Experimento	RMSE	U de Theil
Vale	0,025	1,1
Vale + Ibovespa	0,022	1,02
Vale + Ibov + Sent. #1	0,022	1,02
Vale + Ibov + Sent. #2	**0,021**	**0,98**
Vale + Ibov + Sent. #3	0,022	1,15
Vale + Ibov. + SELIC	0,030	2,05
Vale + Ibov + Dólar	0,063	1,25
Vale + Ibov. + D. Jones	0,024	1,04
JSHF	**0,012**	**0,97**
Usiminas + Ibov. + Sent. #2	**0,016**	**0,98**

V. Conclusões e Trabalhos Futuros

Neste trabalho foi proposto um modelo de *machine learning* capaz de prever o valor de preço médio da ação, para auxiliar investidores na sua tomada de decisão. Uma das ações escolhidas foi a da VALE3 que, no período em análise, teve um comportamento bastante instável devido ao desastre da barragem de Mariana-MG, fazendo com que a rede tivesse mais dificuldade em aprender e também prever os valores. Isso nos permite inferir que, em casos de ações mais estáveis, o resultado seria mais assertivo. Foram utilizadas também as ações das empresas JHSF e Usiminas para validar a hipótese de que todo e qualquer outro ativo que tenha as mesmas informações pode ser utilizado na arquitetura de modelo proposto, necessitando apenas de um novo treinamento.

Os sentimentos extraídos dos *tweets* se mostraram bastante úteis em quase toda a predição porém, em dados momentos, a predição não foi tão acurada como nos outros pontos, podendo levar a uma decisão que não seria tão interessante, pensando em um cenário real. Este fato pode ser em decorrência de não haver *tweets* suficientes ou, mesmo quando os há em números, não há relevância para todos os dias da série histórica, fazendo assim com que a rede possa desconsiderar tais informações em determinados pontos da predição. Um fator limitante da quantidade de *tweets* não ter sido a desejada foi que a API tinha limitações quanto ao número mensal de requisições e limite de conteúdo por requisição.

Como trabalhos futuros, pretende-se analisar *tweets* relacionados a iniciativas ambientais, sociais e de governança (ESG - *Environmental, Social and Governance*) tomadas pela Vale e por outras empresas que passaram por eventos que causaram grande perda no valor das ações, e a relação do sentimento provocado por essas iniciativas com a retomada do valor das ações. Isso porque aspectos não-financeiros, que podem ser quantificados pela análise de sentimento, influenciam no preço das ações.

Agradecimentos

Os autores deste trabalho agradecem os apoios dos projetos: 309737/2021-4-CNPq, 2021-WMR44-FAPES, 85232785-CNPq/FAPES (Edital 22/2018 - Programa Primeiros Projetos).

REFERÊNCIAS

[1] H. Lastres and S. Albagli, *Informação e globalização na era do conhecimento*. Editora Elsevier, 1999. [Online]. Available: https://books.google.com.br/books?id=ZZXitgAACAAJ

[2] E. Schöneburg, "Stock price prediction using neural networks: A project report," *Neurocomputing*, vol. 2, no. 1, pp. 17–27, 1990.

[3] M. L.Lima, T. P. Nascimento, S. Labidi, N. S. Timbo, M. V. L. Batista, G. N. Neto, E. A. M. Costa, and S. R. S. Sousa, "Using Sentiment Analysis for Stock Exchange Prediction," *International Journal of Artificial Intelligence & Applications*, vol. 7, no. 1, pp. 59–67, Jan. 2016. [Online]. Available: http://www.aircconline.com/ijaia/V7N1/7116ijaia06.pdf

[4] S. Borovkova and I. Tsiamas, "An ensemble of lstm neural networks for high-frequency stock market classification," *Journal of Forecasting*, vol. 38, no. 6, pp. 600–619, 2019.

[5] Y. Liu, Z. Qin, P. Li, and T. Wan, "Stock volatility prediction using recurrent neural networks with sentiment analysis," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2017, pp. 192–201.

[6] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, Mar. 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S187775031100007X

[7] S. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of SemEval 2013*. Association for Computational Linguistics, 2013, pp. 321–327. [Online]. Available: https://aclanthology.org/S13-2053

[8] M. Araújo, P. Gonçalves, M. Cha, and F. Benevenuto, "ifeel: a system that compares and combines sentiment analysis methods," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 75–78.

[9] H. R. O. Rocha and M. Macedo, "Previsão do preço de ações usando redes neurais," in *Congresso USP de Iniciação Científica em Contabilidade*, 2011.

[10] C.-R. Ko and H.-T. Chang, "Lstm-based sentiment analysis for stock price forecast," *PeerJ Computer Science*, vol. 7, p. e408, 2021.

[11] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–24, 2020.

[12] Y. Jiang, Y. Zuo, and Y. Yang, "Machine learning and artificial intelligence applied in the research of the emotions impact on forecasting," *Journal of Physics: Conference Series*, vol. 1982, no. 1, p. 012052, jul 2021. [Online]. Available: https://doi.org/10.1088/1742-6596/1982/1/012052

[13] B. A. Januário, A. E. d. O. Carosia, A. E. A. da Silva, and G. P. Coelho, "Sentiment analysis applied to news from the brazilian stock market," *IEEE Latin America Transactions*, vol. 20, no. 3, pp. 512–518, 2021.

[14] F. Audrino, F. Sigrist, and D. Ballinari, "The impact of sentiment and attention measures on stock market volatility," *International Journal of Forecasting*, vol. 36, no. 2, pp. 334–357, 2020.

[15] I. Botunac, A. Panjkota, and M. Matetic, "The importance of time series data filtering for predicting the direction of stock market movement using neural networks," in *Proceedings of the 30th DAAAM International Symposium*, 2019.

[16] A. GUMUS and C. O. SAKAR, "Stock market prediction by combining stock price information and sentiment analysis," *International Journal of Advances in Engineering and Pure Sciences*, vol. 33, no. 1, pp. 18–27, 2021.

[17] D. P. Gandhmal and K. Kumar, "Systematic analysis and review of stock market prediction techniques," *Computer Science Review*, vol. 34, p. 100190, 2019.

[18] T. Swathi, N. Kasiviswanath, and A. A. Rao, "An optimal deep learning-based lstm for stock price prediction using twitter sentiment analysis," *Applied Intelligence*, Mar 2022. [Online]. Available: https://doi.org/10.1007/s10489-022-03175-2

[19] O. Brito, *Mercado financeiro*. Saraiva Educação SA, 2019.

[20] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, ser. Studies in computational intelligence. Berlin: Springer, 2012. [Online]. Available: https://cds.cern.ch/record/1503877

[21] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 855–868, 2008.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] C. Olah, "Understanding lstm networks," 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[24] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008. [Online]. Available: http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html

[25] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, I. C. Society, Ed., 2010.

[26] H. Theil, "The development of international inequality 1960–1985," *Journal of Econometrics*, vol. 42, no. 1, pp. 145–155, 1989.

Gabriel Moura Vargas Gabriel Moura Vargas possui bacharelado em Ciência da Computação pela Universidade Federal do Espírito Santo (2021) e técnico em informática pela Escola Técnica Juscelino Kubitschek (2013). Atualmente é cientista de dados na empresa Minds Digital Informática e seus interesses de pesquisa incluem deep learning, machine learning, análise de dados e mercado financeiro.

Leonardo José Silvestre possui Doutorado em Engenharia Elétrica pela UFMG (2015), Mestrado em Informática pela UFES (2005) e Graduação em Ciência da Computação pela UFV (2003). Atualmente é professor Adjunto IV do Departamento de Computação e Eletrônica do CEUNES/UFES, e seus interesses de pesquisa incluem redes neurais, deep learning e suas aplicações em saúde, smartgrids e agricultura.

Luís Otávio Rigo Júnior possui doutorado em Engenharia de Sistemas e Computação pela UFRJ (2011). Atualmente é professor associado do Departamento de Computação e Eletrônica (DCEL), UFES. Atua na área de Aprendizado de Máquina, desenvolvendo soluções para problemas em energia, saúde e agricultura, bem como soluções de ensino em inteligência artificial.

Helder Roberto de Oliveira Rocha possui Doutorado e Mestrado em Computação Cientifica e Sistemas de Potência - UFF, Bacharelado em Administração - UFRRJ e Graduação em Engenharia Elétrica - UFF. Foi professor no Instituto Federal do Espírito Santo. Ocupa atualmente o cargo de Professor Classe C - Adjunto III do Departamento de Engenharia Elétrica da UFES. É bolsista de Produtividade PQ2.