Usando aprendizado de máquina para predizer sinal do movimento de ativos com auxílio de notícias

1st Stéfano A. S. Spindola
Time de Ciências de Dados
Itaú Unibanco
São Paulo – SP , Brasil
stefano.spindola@itau-unibanco.com.br

2nd Hitoshi Nagano
Tecnologia e Ciências de Dados
FGV EAESP
São Paulo – SP , Brasil
hitoshi.nagano@fgv.br

***Abstract*—Decisions made by investors are influenced by news, changing the state of the market. This makes them an important source of data for financial forecasts. This paper studies how financial forecasting results can improve machine learning models when news articles for a particular asset or a set of daily items are used simultaneously with historical asset price data. Four subsampled data are used from two sources, while recurrent neural network and emsemble models are used. Experimental results show that the simultaneous use of a news aggregate per day improves forecast performance compared to methods based on a smaller number of news categories.**

***Index Terms*—news, market, machine learning**

Resumo - As decisões tomadas por investidores são influenciadas por notícias que alteram o estado do mercado. Isso os torna uma importante fonte de dados para previsões financeiras. Este trabalho de pesquisa estuda como os resultados da previsão financeira podem melhorar com os modelos de aprendizado de maquina (machine learning) quando artigos de notícias para um determinado ativo ou um conjunto de artigos diários são usados simultaneamente com dados histórico de preço dos ativos. São utilzados quatro sub amostra de dados de duas fontes, enquanto algoritmos de rede neurais recorrentes e emsemble são utilizadas. Os resultados experimentais mostram que o uso simultâneo de um agregado de notícias por dia melhora o desempenho da previsão em comparação com métodos baseados em um número menor de categoria de notícias.

Palavras Chaves - notícias, mercado, aprendizado de máquina

I. Introdução

Movimentos de preços de ações são impulsionados por publicações de notícias financeiras. As decisões dos investidores são tomadas com base nas informações disponíveis, considerando como um novo dado influenciará o mercado. Artigos de notícias incorporam informações sobre os fundamentos de uma empresa, as atividades nas quais uma empresa está envolvida e as expectativas de outros participantes do mercado sobre futuras mudanças de preços [1], [2]. Uma enorme quantidade de informação textual é fornecida por aplicações de negociação em tempo real com uma velocidade de transmissão muito grande [3]. Os pesquisadores têm trabalhado em estruturas automatizadas que analisam grandes quantidades de artigos de notícias financeiras, extraem informações relevantes e as utilizam para previsões financeiras [4]. Os métodos existentes de mineração de dados foram empregados e expandidos para estudar como os itens de notícias afetam o preço das ações [5], [6]. Este trabalho de pesquisa estuda como o uso simultâneo de artigos de notícias financeiras podem dar uma vantagem nas previsões. Para atingir esse objetivo, nós utilizamos a API Funcional em Keras, viabilizando o uso de redes neurais recorrentes como LSTM (Long Short Term Memory) e a funcionalidade de multiplas entradas de dados para treinamento dos modelos. A caráter de comparação de modelos, foram utilizados emsemble de modelos como Random Forest Classifier, SVC e outros. As redes LSTM têm uma catacteristica peculiar para o caso de estudo, que é sua capacidade de memória de curto prazo e a hipótese a ser explorada aqui é que esse recurso pode apresentar ganhos em termos de resultados quando comparado a outras abordagens mais tradicionais como por exemplo os modelos ensemble que também serão utilizadas neste estudo. Dados históricos de preços de ativos e dados de noticias alimentaram os modelos simultaneamente com o objetivo de prever o sinal de positivo ou negativo do valor de retorno daqui dez dias das observações atuais. Por fim métricas adequadas de desempenhos dos modelos foram empregadas, servindo de parâmetro de avaliação de cada extratégia adotada. Os experimentos mostram que uma tentativa de agregar notícias por dia e fazer a predição de cada observação de ativos melhora as previsões, demonstrando um desempenho promissor quando comparado com abordagens baseadas em um único subconjunto de notícias. O restante desta pesquisa está organizado da seguinte forma. A seção II fornece uma revisão da literatura relacionada ao caso de estudo. A seção III especifica dados brutos, descreve o préprocessamento de dados e as abordagens de aprendizado de máquina usadas e descreve as métricas de desempenho usadas para análise. Seção IV discute resultados experimentais. A seção V contém conclusões e incentivos para trabalhos futuros.

II. Revisão bibliográfica

Os sistemas de negociação fornecem uma enorme quantidade de dados textuais para os comerciantes do mercado de capitais. Anúncios oficiais, recomendações de analistas, jornais financeiros, fóruns de discussão e feeds de notícias de serviços de notícias são exemplos de informações disponíveis para investidores [5]. Wüthrich et al. [7] fizeram a primeira tentativa de usar informações textuais para previsão do mercado de ações. Um dicionário de termos obtidos de um especialista foi utilizado para atribuir ponderações a recursos e gerar regras probabilísticas para prever as variações diárias de preço de cinco índices de ações. Uma estratégia de negociação baseada nas previsões do sistema demonstrou que os retornos positivos podem ser obtidos usando notícias financeiras. Lavrenko et al. [8] desenvolveram o sistema Analyst, que incluiu modelos de linguagem, utilizou a série temporal de preços e classificou as notícias recebidas. Os autores mostraram que os lucros podem ser produzidos usando o sistema projetado. Gidofalvi e Elkan [9] criaram um sistema para prever movimentos de preços de curto prazo. Os artigos de notícias foram alinhados, classificados por meio de regressão linear em relação ao índice NASDAQ, em seguida, atribuídos com um rótulo "para cima", "para baixo" ou "inalterado". O autor concluiu que o comportamento das ações está fortemente correlacionado com o conteúdo de uma reportagem de 20 minutos antes de 20 minutos após a publicação. Chan [10] examinou os retornos mensais usando manchetes sobre empresas específicas e descobriu que as publicações de más notícias causam fortes desvios negativos no mercado. Kloptchenko et al. [11] concentraram-se em relatórios oficiais da empresa e confirmaram sua capacidade de indicar o desempenho futuro da empresa. Por exemplo, uma mudança no estilo de um relatório pode indicar uma mudança significativa na produtividade da empresa. As relações entre os retornos médios das empresas e sua cobertura da mídia foram examinadas por Fang e Peress em [12]. Os autores concluíram que as ações com alta cobertura tiveram um desempenho significativamente inferior comparado com ações não apresentadas na mídia. Garcia [13] investigou a relação entre sentimentos de artigos no New York Times e retornos de ações e concluiu que o conteúdo de notícias ajuda a prever retornos de ações e que os sentimentos dos investidores têm um efeito notável durante as recessões. Tetlock [14] examinou as interações entre o conteúdo dos artigos diários publicados pelo Wall Street Journal e as ações. Os resultados mostraram que notícias altamente pessimistas causam uma queda nos preços de mercado e aumentam notavelmente o volume de negociações.

III. Projeto de Pesquisa

Esta seção fornece detalhes sobre o projeto de pesquisa do sistema preditivo baseado em notícias. Ele explica como as notícias foram especificadas, descreve dados textuais brutos e discute o pré-processamento de dados, abordagens de aprendizado de máquina e métricas de desempenho usadas para avaliação dos modelos.

A. Notícias

Os artigos de notícias financeiras publicadas sobre ativos foram obtidas através da coleta de dados por meio de crawlers e outra de uma fonte privada também do kaggle, ambas possuem artigos durante um periodo de 9 anos, de 1° de Janeiro de 2007 a 30 de Dezembro de 2016. O provedor de notícias para ambas as bases é a Reuters News que mostraram suficiente cobertura da imprensa sobre ações que compõem o índice de mercado S P 500 da base publica e os 3780 ativos do estoque da base privada. O número total dos artigos de notícias coletados de ambas bases é de aproximadamente 9.328.750 de artigos. Para cada artigo de ambos as bases, as seguintes informações foram armazenadas: ano, mês, dia e headline (título) de notícias. Para cada artigo, uma data correspondente de publicação é construída a partir do dia, mês e ano. No caso da base privada foi armazenados também mais dados que são interessantes, porque traz informações importantes do corpo da notícia, como o grau de relevância do artigo para as empresas, o número total de sentenças no corpo de notícia para cada item, o número total de tokens lexicais (palavras e pontuação) na notícia, a classe de sentimento predominante para este item de notícias em relação ao ativo (a classe indicada é aquela com maior probabilidade), sentimento negativo em probabilidade, sentimento neutro, sentimento positivo, o número de tokens lexicais nas seções do texto do item que são considerados relevantes para o ativo. Mas para haver reprodutividade deste trabalho, selecionamos apenas as informações que contem também na base pública. Como resultado, dois subconjuntos de dados foram formados através da base de dados públicos e dois subconjunto de dados foram formados da base de dados privada das seguintes formas, ambas possuem subconjuntos de dados de todas as headlines em uma série de notícias e todas as headlines referente a um dia especifico agregados em um dia unico respectivamente conforme pode ser observado nas Tabelas I e II .

Tabela I
Estrutura dos subconjuntos gerados da Base Pública

Conjunto de Dados	Base Pública	
Subconjunto 1	Data	Headline de notícias
Subconjunto 2	Data	Headline de notícias agregadas por dia

Tabela II
Estrutura dos subconjuntos gerados da Base Pivada

Conjunto de Dados	Base Privada	
Subconjunto 1	Data	Headline de notícias
Subconjunto 2	Data	Headline de notícias agregadas por dia

No sistema de previsão, as previsões foram feitas somente após as datas em que pelo menos um artigo foi publicado . Vários pontos de dados foram definidos nesse estágio para cada ação. Uma única instância de dados corresponde a uma data em que um artigo específico de um ativo foi publicado. Ao formar outros subconjuntos de dados, os artigos relevantes

para um item de ativo foram agregadas apenas para este dia correspondente ao ativo.

B. Dados históricos de preços

Os preços históricos foram usados para selecionar os recursos mais expressivos dos artigos de notícias e para rotular as instâncias de dados. Séries temporais de preços dos mesmos perfis do kaggle já mencionados, portanto dados públicos e privados. Para as bases públicas os recursos foram selecionados com base no feedback do mercado definido como um movimento do preço das ações no próximo dia de negociação após o dia da publicação. O movimento do preço foi definido como a diferença entre os preços das ações em aberto e no próximo dia de negociação. Para a base privada temos os valores de abertura, fechamento, volume de negociações dos ativos. Um problema de classificação de duas classes foi considerado neste trabalho de pesquisa. Os rótulos "sobe" ou "desce" que corresponderiam a um aumento ou diminuição no preço dos ativos nos próximos dez dias respectivamente, foram atribuídos a cada instância de dados.

C. Pré-processamento de dados textuais

Como etapa preliminar final, o conjunto de dados de artigos é convertido em um formato apropriado para aprendizado de máquina. Aqui, cada artigo de notícias passa por um tokenizador de texto, que permite que vetorize um corpus de texto, transformando cada texto em uma sequência de inteiros (cada inteiro sendo o índice de um token em um dicionário) ou em um vetor em que o coeficiente de cada token poder ser binário, com base na contagem de palavras, baseado em TF-IDF que são computados para representar cada característica. Quando um recurso não ocorre em um artigo, seu valor TF * IDF é igual a zero. Portanto, o resultado é uma matriz esparsa, em que cada valor é igual ao valor correspondente de TF * IDF. Depois que o pré-processamento textual for concluído, os rótulos "sobe" ou "desce"serão atribuídos aos pontos de dados. Está etapa foi utilizada para obtermos os dados de entrada para treinamento da rede neural recorrente LSTM. O mesmo procedimento é nessesário para treinamento do ensemble, mas neste caso específico Countvectorize que também transforma os dados em uma matriz esparsa, performou melhor que a técnica anterior.

D. Técnicas de aprendizado de máquina

O conjunto de dados foi dividido em conjuntos de dados de treinamento, validação e teste em ordem cronológica. Primeiros 90% dos 500.000 pontos de dados selecionados do período de 1º de Janeiro de 2016 à 30 de Maio de 2018 foram usados para treinar os modelos. Os 10% subsequentes dos pontos de dados foram usados para validação, uma fase necessária para ajustar os parâmetros do modelo. Para a classificação com LSTM, utilizamos um modelo de múltiplas entradas, alimentando a rede com dados de preços em uma camada densa e os dados de notícias em uma camada embedding seguido de uma camada com LSTM, feito isso concatenamos as saídas de ambas as camadas e alimentando um camada Densa com função de ativação sigmoid para nos dar a classificação das classes binárias. Na Figura 1 poderá ver como foi o processo de aprendizado.

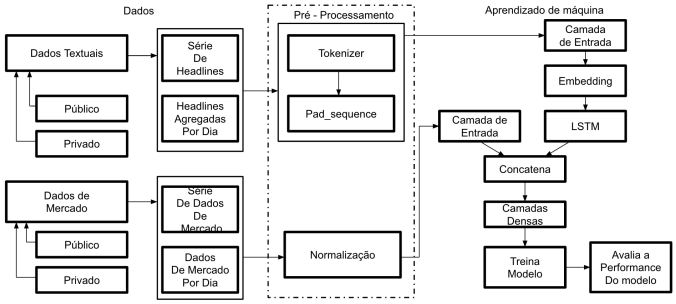

Figura 1. Arquitetura de aprendizado do modelo LSTM

Para a classificação com ensemble, foram utilizados vários modelos (Random Forest Classifier, Regressão Logistica, BernoulliNB, SVC, Kneighbors Classifier) sobre os artigos de notícias e utilizando a saída do treinamento de texto como atributo para os dados de mercado e treinando com as características dos dois dados, para explorar o poder conjunto desses dados para a predição. Os 10% restantes das instâncias de dados foram utilizados para testar o sistema preditivo desenvolvido. A precisão foi usada para medir o desempenho do modelo com diferentes configurações de parâmetros durante a fase de validação. Na Figura 2 poderá ver como foi o processo de aprendizado.

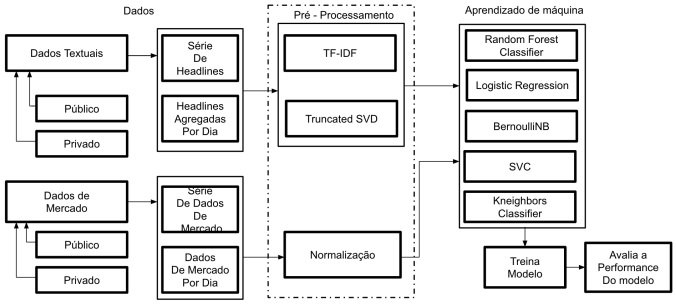

Figura 2. Arquitetura de aprendizado do ensemble de modelos

E. Medidas de Desempenho

Para cada um dos conjuntos de dados, o desempenho preditivo das técnicas de aprendizado de máquina foram medidas usando a acurácia, precisão e revocação de previsão.

- Acurácia retorna o quão frequente o classificador está correto e pode ser expressa por:

$$Acurácia = \frac{VP + VN}{Total} \tag{1}$$

VP = Verdadeiros Positivos; VN = Verdadeiros Negativos

- Precisão retorna de uma determinada classe quantas efetivamente classifiquei corretamente e pode ser expressa por:

$$Precisão = \frac{VP}{VP + FP} \tag{2}$$

Tabela III
AVALIAÇÃO DOS RESULTADOS OBTIDOS PELOS MODELOS DE MACHINE LEARNING, APLICADOS NA BASE PÚBLICA

Técnicas de Machine Learning	Base Pública					
	Subconjunto 1			Subconjunto 2		
	Acurácia	Precisão	Revocação	Acurácia	Precisão	Revocação
LSTM	67,6%	67,5%	64,2%	65,2%	60,2%	59,0%
Random Forest Classifier	64,1%	32,4%	62,1%	63,5%	30,2%	61,5%
Logistic Regression	64,3%	32,3%	63,7%	62,9%	30,0%	60,3%
BernoulliNB	54,5%	32,4%	53,0%	52,9%	30,7%	60,1%
SVC	63,9%	32,2%	63,0%	62,5%	30,6%	60,2%
Kneighbors Classifier	62,5%	49,1%	63,5%	52,7%	50,3%	60,0%

Tabela IV
AVALIAÇÃO DOS RESULTADOS OBTIDOS PELOS MODELOS DE MACHINE LEARNING, APLICADOS NA BASE PRIVADA

Técnicas de Machine Learning	Base Privada					
	Subconjunto 1			Subconjunto 2		
	Acurácia	Precisão	Revocação	Acurácia	Precisão	Revocação
LSTM	54,2%	67,5%	49,5%	53,1%	60,1%	50,0%
Random Forest Classifier	54,9%	34,4%	53,9%	53,4%	31,7%	61,7%
Logistic Regression	55,5%	33,4%	55,0%	53,8%	33,6%	60,5%
BernoulliNB	53,0%	32,4%	48,7%	52,9%	31,4%	60,9%
SVC	53,3%	32,2%	52,5%	53,3%	33,7%	60,1%
Kneighbors Classifier	51,1%	49,1%	52,0%	50,9%	40,2%	50,0%

FP = Falsos Positivos

- Revocação retorna o quão frequente você classifica como uma determinada classe, em outras palavras é a frequência em que seu classificador encontra os exemplos de uma classe e pode ser expressa por:

$$Revocação = \frac{VP}{VP + FN} \tag{3}$$

FN = Falsos Negativos

IV. RESULTADOS EXPERIMENTAIS

Esta seção mostra os resultados experimentais produzidos pelo sistema de previsão baseado em notícias. As acurácias, precisões e revocações apresentados nas tabelas abaixo foram obtivas dos 4 subconjuntos analisados.

A. LSTM

Esta subseção analisa o desempenho dos classificadores LSTM. A Tabela III e Tabela IV descrevem os resultados de predição obtidos utilizando a mesma arquitetura de rede para os 4 subconjuntos de dados. A maior precisão foi obtida para o subconjunto 1 da base pública, no qual foram concatenadas todas as notícias que corresponde a um determinado dia e usado para realizar a predição junto com o poder de predição dos preços dos ativos, portanto o modelo obteve uma performance melhor para o subconjunto 1 da base pública.

B. Ensemble

Para os modelos emsemble, podemos observar que não foram melhores em todos os subconjuntos testados e com uma baixa acurácia e precisão, mas podemos observar retultados melhores nos testes na base pública, principalmente no subconjunto 1 de dados, mas não é melhor que a LSTM para o ambos os subconjuntos.

V. CONCLUSÃO E TRABALHOS FUTUROS

Este estudo de pesquisa explora se o uso simultâneo de notícias financeiras com dados de preço de ativos podem fornecer uma vantagem no sistema de previsão financeira baseado em notícias. Quatro conjuntos de dados foram utilizados com artigos de notícias e preço de ativos. Cada conjunto foi pré-processado de forma a utilizarmos os modelos propostos e gerar as predições. A abordagem com LSTM apresentou melhores resultados de acurácia, precisão e revocação de previsão para os subconjuntos de dados 1 da base pública. Os métodos ensemble demonstraram em geral um desempenho não tão bom quanto a LSTM, mas entre eles destaca-se o modelo Logistic Regression que em geral performou muito bem só perdendo em desempenho para a LSTM nos subconjuntos de dados 1 para a base pública. Esses resultados indicam que utilizar itens de notícias com base de preço de ativos permitem que o sistema aprenda e utilize mais informações sobre o comportamento futuro dos preços, o que dá uma vantagem para previsões mais precisas. Os resultados alcançados são promissores considerando a proposta deste trabalho e com a continuidade da pesquisa, adicionando técnicas mais robustas sobre o conjunto de notícias e sobre os dados históricos dos ativos, acredito na viabilidade de utilizar a combinação dos dois dados de natureza distinta para os modelos de aprendizado de máquina. A introdução de fontes de dados adicionais, como sentimento de notícias e outras características das notícias originais podem melhorar o desempenho dos modelos preditivos. Outras possíveis direções de trabalho futuro envolvem a melhor configuração dos modelos de classificação e adicionar sentimentos dos itens de notícias.

REFERÊNCIAS

[1] M. Hagenau, M. Liebmann, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context- capturing features," Decision Support Systems, vol. 55, no. 3, pp. 685–697, 2013.

[2] X. Zhao, J. Yang, L. Zhao, and Q. Li, "The impact of news on stock market: Quantifying the content of internet-based financial news," in Proc. of the 11th Intl DSI 16th APDSI Joint meeting, 2011, pp. 12–16

[3] S. S. Groth and J. Muntermann, "An intraday market risk management approach based on textual analysis," Decision Support Systems, vol. 50, no. 4, pp. 680–691, 2011.

[4] G. Mitra and L. Mitra, The handbook of news analytics in finance. Wiley-Finance, 2011.

[5] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news," ACM Transactions on Information Systems, vol. 27, no. 2, pp. 1–19, 2009.

[6] M. Mittermayer and G. Knolmayer, "Text mining systems for market response to news: A survey," vol. 41, no. 184. University of Bern, 2006.

[7] B. Wiithrich, D. Permunetilleke, and S. Leung, "Daily prediction of major stock indices from textual www data," in Proc. of the 4th Intl Conference on Knowledge Discovery and Data Mining, 1998.

[8] V. Lavrenko, M. Schmill, and D. Lawrie, "Mining of concurrent text and time series," in Proc. of the 6h ACM Intl Conference on Knowledge Discovery and Data Mining, 2000.

[9] G. Gidófalvi and C. Elkan, "Using news articles to predict stock price movements," Department of Computer Science and Engineering, University of California. 2001.

[10] W. S. Chan, "Stock price reaction to news and no-news: drift and reversal after headlines," Journal of Financial Economics, vol. 70, no. 2, pp. 223–260, 2003.

[11] A. Kloptchenko, T. Eklund, B. Back, J. Karlsson, H. Vanharanta, and A. Visa, "Combining data and text mining techniques for analysing financial reports," Intl Journal of Intelligent Systems in Accounting and Finance Management, vol. 12, no. 1, pp. 29 – 41, 2004.

[12] L. Fang and J. Peress, "Media Coverage and the Cross-section of Stock Returns," The Journal of Finance, vol. LXIV, no. 5, pp. 2023–2052, 2009.]

[13] D. Garcia, "Sentiment during recessions," The Journal of Finance, vol. 68, no. 3, pp. 1267–1300, 2013.

[14] P. C. Tetlock, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," Journal of Finance, vol. 62, no. 3, pp. 1139–1168, 2007.