Previsão de Ações da JBS

Nicola Pedulla, Ricardo Barioni

December 10, 2017

Abstract

A análise de dados no contexto de aplicações na bolsa de valores é muito importante, pois dá às pessoas e empresas que praticam esse tipo de atividade informações mais factíveis a respeito de que tipos de investimentos são interessantes de se realizar, em função da situação do mercado. Para tal, foi realizada uma análise de comportamento do preço de ações, mais especificamente da empresa JBS, em função de informações contidas na internet, estas mais especificamente provenientes do Twitter e do Google Trends, com o objetivo de auxiliar na previsão do valor dessas ações.

Keywords: Bolsa de Valores, Twitter, Google Trends

1. Introdução

Como se sabe, no contexto de aplicação de investimentos na bolsa de valores, é necessário diversos tipos de informações e conhecimentos. Primeiramente, é essencial que o investidor tenha conhecimento da situação das instituições a serem investidas perante ao mercado que o rodeia; para isso, o investidor precisa estar em sincronia com as informações mais recentes (estas provenientes de noticiários, revistas ou até mesmo informações confidenciais o qual o investidor tenha conhecimento) as quais rodeiam (e possam influenciar) determinada entidade a se investir.

Não só isso, como um investidor interpreta essas informações? Isso aí está relacionado a um outro diferencial, o qual é entender os efeitos colaterais desses acontecimentos. Além disso, esses efeitos colaterais podem variar no tempo, ou seja, eles podem acontecer imediatamente ou a longo prazo, em função do acontecimento. Porém, o mercado de ações é algo bem complexo, uma vez que existem diversos fatores os quais podem influenciar o preço de ações de uma entidade.

Com o avanço da tecnologia da informação, realizar a análise de alguns desses fatores tornou-se mais fácil; a inclusão tecnológica foi responsável por contemplar diversos indivíduos com a capacidade de expressarem sua opinião, bem como a possibilidade de buscar por informações livremente. Com esses dois tipos de informações, já é possível circundar alguns fatores os quais estão atrelados à análise de preços da bolsa de valores.

Focando na abordagem realizada por esse trabalho, usou-se informações relacionadas a tweets (provenientes da rede social Twitter) os quais possuem alguma relação com algum acontecimento relevante à empresa JBS, além de informações extraídas da plataforma Google Trends, sendo também relacionadas à empresa JBS. Com essas informações em conjunto, avaliou-se o comportamento dos valores das ações da empresa JBS, a fim de analisar se há relações diretas entre esses dados, bem como a força dessas relações.

Para tal, este trabalho ficou subdividido em duas etapas: a coleta dos dados provenientes da internet, e efetivamente a análise desses dados.

2. Análise

2.1. Coleta dos dados

A coleta dos dados consistiu-se em duas partes: a extração de tweets e dados referentes ao Google Trends.

Primeiramente, levamos em consideração os tweets relacionados à JBS, nos quais também são referentes a algum acontecimento potencialmente relevante à empresa. Para tal, escolhemos o evento de delação da JBS, o qual aconteceu no dia 17 de maio de 2017. Logo, a coleta de tweets foi efetuada para tweets entre seis meses antes e depois do ocorrido.

Já para os dados do Google Trends, foram coletados informações do quão "em alta" está o assunto "JBS", dentre os dias 26 de fevereiro de 2017 e 2 de outubro de 2017.

2.1.1. Twitter

Primeiramente, utilizou-se a api "GetOldTweets-python", com o objetivo de coletar tweets de datas menos recentes, uma vez que o uso da api "tweepy" limita a coleta de tweets para os últimos 15 dias. Com isso, coletou-se 1000 tweets por dia (para todos o intervalo de tempo descrito anteriormente, caso houvessem 1000 tweets para cada dia) provenientes da busca "JBS". Com isso, um total de 243002 tweets foram coletados.

Porém, vale considerar que muitos desses tweets podem não estar relacionados com o escândalo de delações da JBS. Para isso, criamos um classificador capaz de avaliar se um tweet está (ou não) relacionado ao tema, levando em consideração a sua estrutura.

Primeiramente, escolhemos um conjunto de 500 tweets vetorizados, os quais classificamos como "relacionados à delação da JBS" ou "não relacionados à delação da JBS". Com isso, realizamos a vetorização dos tweets pelo método "Bag of Words", no qual um tweet possui uma representação relacionada à quantidade de palavras existentes no vocabulário de tweets.

Tendo esses 500 tweets classificados em mãos, realizou-se o processo de Stratified K-fold; tal conjunto é dividido em dois subconjuntos complementares, no qual um é denominado o conjunto de treino, e o outro é o conjunto de teste. Com isso, gera-se um classificador Naive-Bayes baseado no conjunto de treinamento, e a acurácia do classificador é calculada a partir do conjunto de teste. Tal processo foi repetido 100 vezes, e o classificador selecionado é o que obter a maior acurácia no processo.

Com o classificador em mãos, foi realizado a classificação dos 243003 tweets em "relacionado à delação da JBS" ou "não relacionado à delação da JBS". Com essa classificação, obteve-se o resultado da quantidade de tweets por dia, bem como a quantidade de tweets relacionados (ou não relacionados) ao evento, por dia.

Para acoplar as informações dos tweets, ao processo de análise, a janela de tweets foi limitada ao intervalo de dias descrito entre 26 de fevereiro de 2017 e 2 de outubro de 2017. Além disso, as informações diárias foram compostas em informações semanais.

2.1.2. Google Trends

Para obter os dados do Google Trends, utilizamos uma api fornecida no Github, chamada "pytrends".

2.2. Pré-processamento dos dados

No pré-processamento dos dados, foi escolhido dividir o conjunto em semanas para tentar facilitar a análise. Também criamos outras colunas: "Previous searchs" e "Previous tweets", que são os valores de pesquisas e de tweets feitos na semana anterior e que serão utilizado para prever o valor da ação atual. No caso da variável "Close" (preço da ação no dia), o valor se tornou a média dos valores durante a semana.

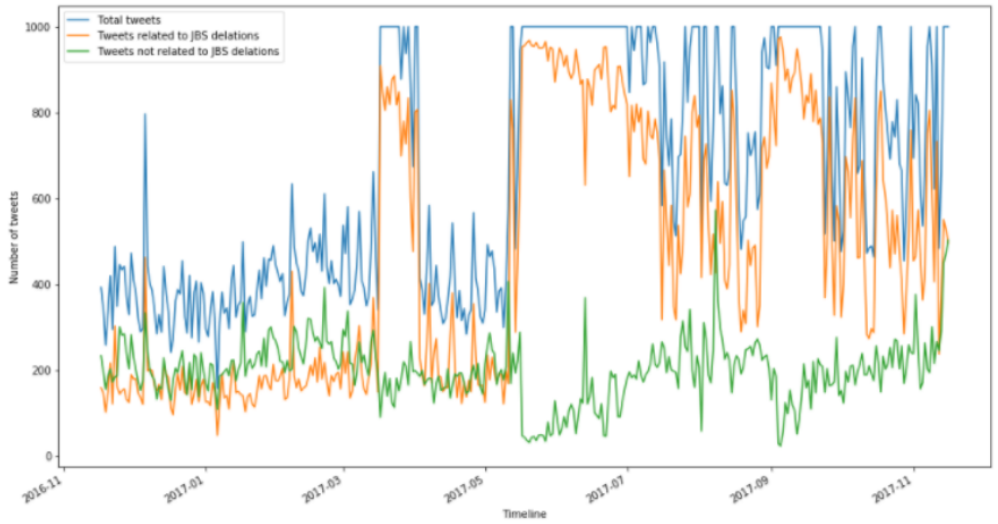

Figure 1: Gráfico da quantidade de tweets ao longo do tempo. A linha azul representa a quantidade total de tweets por dia, a linha laranja representa os tweets relativos à delação da JBS, e a linha verde representa os tweets não relativos à delação da JBS.

	Close	Volume	searchs	tweets	Previous searchs	Previous Tweets	Relative Change searchs	Relative Change tweets	Relative Change Close
2017-01-01	11.424	5937540.0	4	863	NaN	NaN	NaN	NaN	NaN
2017-01-08	11.892	10434640.0	5	1132	4.0	863.0	NaN	NaN	0.040966
2017-01-15	11.830	10532920.0	6	945	5.0	1132.0	0.250000	0.311703	-0.005214
2017-01-22	11.970	7243725.0	5	1146	6.0	945.0	0.200000	-0.165194	0.011834
2017-01-29	12.228	6680480.0	5	1244	5.0	1146.0	-0.166667	0.212698	0.021554

Figure 2: Resultados do pré-processamento realizado.

2.3. Análise

Restringimos a nossa análise no período de 26/02/2017 a 02/10/2017, onde são cobertos os três maiores picos do Google Trends.

Utilizamos duas abordagens para analisar o problema: Mudança relativa, na qual pretendemos analisar se, dado uma variação nas buscas de uma semana para outra, quanto seria variação do preço da ação; e "log transformation", no qual tentamos linearizar os dados para dar uma melhor predição para a regressão linear.

2.3.1. Estimando os dados com regressão linear

Com o Google Trends, a regressão linear não obteve um bom resultado com a variável independente "Log searchs". Ela obteve um coeficiente de

determinação de 0.167368171526 e um RMES de 0.0241384232615. Uma possível causa é a presença de outliers.

Já o Twitter obteve um bom desempenho, com um alto coeficiente de determinação igual a 0.80828719118. Esse modelo consegue explicar mais de 80% dos dados.

2.3.2. Estimando os dados com Random Forest Regressor

O Random Forest teve um resultado pior que a regressão linear. Talvez esse algoritmo não seja ideal para um pequena base de dados.

Com as informações do Google Trends, Random Forest não conseguiu explicar nada dos dados, uma vez que obteve um coeficiente de determinação abaixo de zero.

Com as informações do Twitter, o Random Forest conseguiu explicar bem os dados, obtendo um coeficiente de determinação de 0.787371547718, o que é aproximadamente 80% dos dados.

2.3.3. Estimando os dados com Mudança Relativa

Não obtivemos um bom resultado estimando os dados com mudança relativa. Tanto Random Forest quanto regressão linear obtiveram um coeficiente de determinação negativo para os valores do Google Trends. Na análise dos dados do Twitter, esse modelo teve uma queda de rendimento comparado ao "log transformation", tendo um coeficiente de determinação de 0.14132603756 para regressão linear e de 0.462134946424 para Random Forest.

3. Conclusão

Foi desenvolvida uma análise, na qual comparou-se o contexto dos Tweets relacionados à JBS e informações provenientes do Google Trends com o valor de ações da JBS ao longo do tempo, e foi desenvolvido modelos com o intuito de predizer os futuros valores das ações, em função dessas informações provenientes da internet.

Os resultados são satisfatórios, pois foi possível observar correlações entre essas informações e o valor das ações. Também pode-se observar que essa correlação é mais forte em relação aos dados provenientes do Twitter.

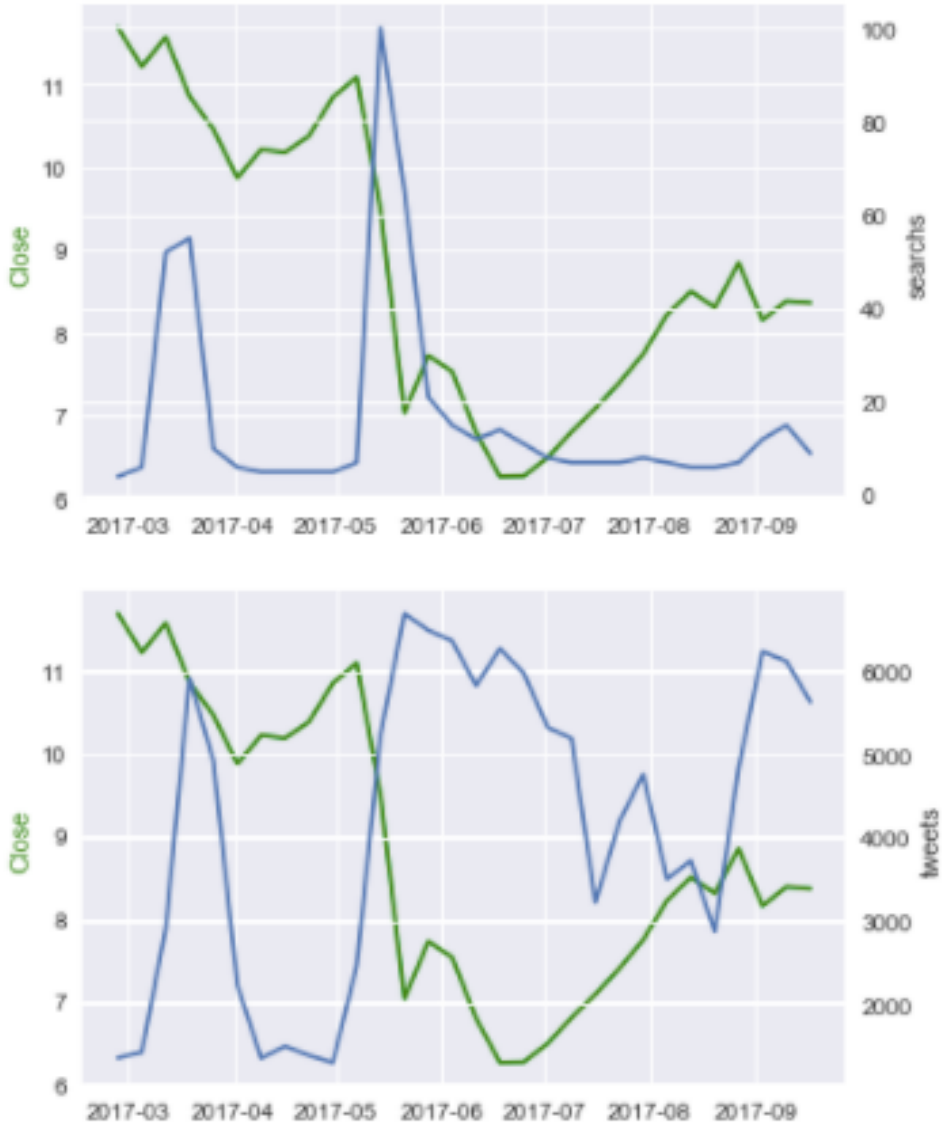

Figure 3: Fazendo uma comparação entre o preço da ação e o valor no Google Trends e entre o preço da ação e os tweets, podemos perceber que para o caso da empresa JBS o preço da ação e as variáveis independentes tem o sentido contrário.

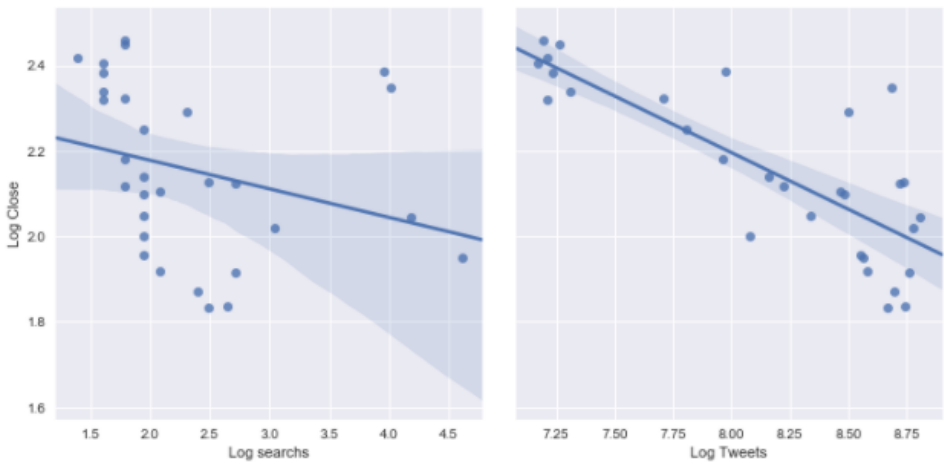

Figure 4: Regressões lineares para o Google Trends e Twitter, pela abordagem "log transformation".

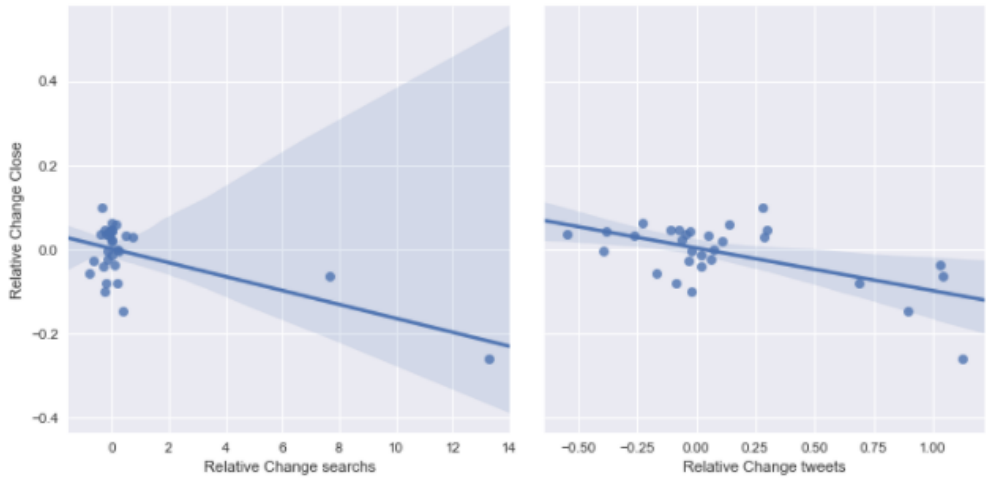

Figure 5: Resultados da aplicação da mudança relativa.