Processamento de Linguagem Natural para predições no mercado financeiro

Thiago Issao Yasunaka (RA103069)[1]

[1]Departamento de Informática, Universidade Estadual de Maringá, Maringá, Brasil
[2]Introdução a Inteligência Artificial, Wagner Igarashi, Trabalho 2

December 7, 2021

1 Introdução

No mercado financeiro existem milhares de formas para investir o seu dinheiro. Uma das formas mais conhecidas para investimentos é por meio do mercado acionário. Segundo Mill [2017], o mercado acionário é o lugar onde pessoas físicas e jurídicas compram e vendem ações. Ações, por sua vez, consiste de pequenos fragmentos de uma determinada empresa, que se somadas, formam o valor de mercado para essa mesma empresa. Por exemplo, uma das empresas mais famosas do Brasil atualmente é a Petrobras, uma empresa de capital aberto que possui ações listadas na bolsa de valores do Brasil, a B3. A Petrobras possui milhões de ações listadas na bolsa, isto faz com que este papel se torne negociável por qualquer pessoa no mundo, de forma pública.

Por se tratar de um mercado de compra e venda que reage em tempo real, as negociações de ações torna o valor de uma empresa extremamente volátil, essa é uma das características que fazem a bolsa de valores ser considerada uma aplicação de renda variável. Quando a demanda por compra de ações é maior do que a demanda por venda, o preço das ações de uma empresa tende a subir. Porém, quando a venda desses papéis tende a ser maior do que a demanda por compra, a tendência para o preço de uma ação é de queda. Essa é uma característica comum em qualquer mercado, conhecido como lei da oferta e da procura.

Um dos principais fatores que fazem uma pessoa ou empresa comprar ou vender ações geralmente é por meio das notícias que são disponibilizadas via internet. Isto significa que uma notícia ruim pode gerar graves consequências no mercado, sendo que o inverso vale para uma notícia boa. Por exemplo, uma notícia de que a Petrobras está envolvida com casos de corrupção, gerando prejuízos para a empresa, é um forte sinal para que investidores não permaneçam mais como acionistas dessa empresa, fazendo com que eles vendam estes papéis. Se a maioria dos investidores pensarem da mesma forma, a lei da oferta e da procura entra em ação e, consequentemente, com uma grande quantidade de ações sendo vendidas no mercado, a tendência do preço dessa ação cair é alta. Agora, supondo que a Petrobras informa à seus acionistas lucros recordes para a companhia, significa que o

mercado provavelmente irá reagir positivamente à esta notícia. E como consequência disso, a busca para comprar ações da Petrobras é maior, fazendo com que o preço da ação suba. Porém esse é o cenário ideal. Como empresas, mercado e pessoas são seres que não são facilmente previsíveis assim como ocorre com as leis da física, nem sempre esse padrão será seguido.

Cientistas da computação, assim como outras ciências, tentam diariamente encontrar soluções para as leis que regem a natureza. Isso não é diferente do tratamento para o mercado acionário. Apesar de não poder ser tratado como um verdade absoluta, um comportamento que pode ser observado em épocas passadas do mercado acionário é que, notícias positivas animam o mercado e notícias negativas colocam medo no mercado. Porém, como dito anteriormente, comportamentos do mercado no passado não pode ser concluído como garantia de que o mesmo comportamento será seguido no futuro, no entanto, um estudo observando esse padrão com as notícias pode ser realizado utilizando algumas técnicas de processamento de linguagem natural. O objetivo desse trabalho será implementar um algoritmo com técnicas de PLN para analisar os sentimentos do mercado de ações por meio de notícias financeiras, e tirar diversas conclusões a partir dessa análise.

2 Fundamentação Teórica

Os seres humanos tem a capacidade de comunicarem entre si por meio de uma linguagem natural. Já os computadores se comunicam por meio de linguagens formais. A princípio, a principal diferença entre os dois tipos de linguagem é a ambiguidade. Linguagens computacionais devem ser precisamente definidos, por exemplo, "print(2 + 2)" é uma linguagem que um interpretador ou compilador entenderia, enquanto que "print(2 +)2" o compilador já acusaria um erro sintático. Já as linguagens naturais são ambíguas, elas podem ter significados diferentes para diferentes contextos. Imagine uma pessoa na praça de uma cidade, ela diz: "Preciso ir ao banco". Essa frase possui diferentes significados, pois essa pessoa poderia tanto ir ao banco como instituição financeira ou como banco se referenciando à uma peça de imobiliário. A partir desso, fazer com que computadores entendam linguagens naturais exige grande esforço de implementação e isso as fazem serem difíceis de trabalhar, deste modo, o que existe na atualidade são aproximações do modelo de língua natural. Segundo Russel and Norvig [2010], existem diversos modelos para tratar linguagens naturais, algumas das principais são: Modelos de caracteres de N-grama, funções de pontuação para RI, algoritmo *PageRank*, extração de informações e etc.

Para fazer a identificação de notícias do mercado financeiro e definir se uma notícia é positiva ou negativa, é necessário encontrar algum modelo que consiga fazer essa análise. Um outro princípio, o classificador de *Naive Bayes* consiste de um modelo probabilístico de aprendizado de máquina que faz essa tarefa de classificação. Este classificador utiliza estatística como ferramenta principal, sendo mais específico, o teorema de *Bayes*. De forma simplificada, segundo Gandhi, o teorema de Bayes consiste em encontrar a probabilidade de um evento A ocorrer, dado que o evento B ocorreu. Observe a figura 1, ela representa a fórmula matemática para encontrar a probabilidade do evento A ocorrer, a partir da probabilidade do evento B.

O classificador de *Naive Bayes* faz uma adaptação ao modelo padrão do teorema de *Bayes*. Considerando que exista um *dataset* (conjunto de dados) disponível e que, para cada informação tenha

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 1: Fórmula básica do teorema de *Bayes*

uma avaliação positiva e negativa, é possível aplicar o teorema de *Bayes* e obter dados relevantes para outras instâncias de modo que ele tenha como base o *dataset* inicial. Existem diversas API's na internet disponível para aplicar o classificador de *Naive Bayes* de forma abstraída, facilitando o seu uso.

```
>>> train = [
...     ('I love this sandwich.', 'pos'),
...     ('this is an amazing place!', 'pos'),
...     ('I feel very good about these beers.', 'pos'),
...     ('this is my best work.', 'pos'),
...     ("what an awesome view", 'pos'),
...     ('I do not like this restaurant', 'neg'),
...     ('I am tired of this stuff.', 'neg'),
...     ("I can't deal with this", 'neg'),
...     ('he is my sworn enemy!', 'neg'),
...     ('my boss is horrible.', 'neg')
... ]
>>> test = [
...     ('the beer was good.', 'pos'),
...     ('I do not enjoy my job', 'neg'),
...     ("I ain't feeling dandy today.", 'neg'),
...     ("I feel amazing!", 'pos'),
...     ('Gary is a friend of mine.', 'pos'),
...     ("I can't believe I'm doing this.", 'neg')
... ]
```

Figure 2: Exemplo de aplicação do classificador de *Naive Bayes*. Imagem retirada de: https://textblob.readthedocs.io/en/dev/classifiers.html

Observe a figura 2, o conjunto de dados nomeado de *train* é processado por um conjunto de operações que aplica o teorema de *Bayes*. Concluindo esse processamento, é possível fazer o processamento de linguagem natural para outras frases que são semelhantes à este *dataset*. Este é um exemplo totalmente abstraído da parte matemática do classificador de *Bayes*, uma biblioteca para a linguagem *Python* chamada *TextBlob*. Para testar uma nova instância para a PLN criada, basta repetir a instrução na figura 3. Além da informação positiva ou negativa, é possível obter por meio da biblioteca *TextBlob* outras informações, como por exemplo, probabilidade da informação ser positiva e probabilidade da informação ser negativa.

```
>>> cl.classify("This is an amazing library!")
'pos'
```

Figure 3: Exemplo de uso do classificador de *Naive Bayes*. Imagem retirada de: https://textblob.readthedocs.io/en/dev/classifiers.html

3 Metodologia

Para a implementação da inteligência artificial, foram utilizadas algumas ferramentas de desenvolvimento que serão descritas na tabela e tópicos abaixo:

- Linguagem de Programação: Python3
- Biblioteca *numpy* para manipulação de matrizes
- Biblioteca *matplotlib* para geração de gráficos
- Biblioteca *csv* para manipulação do *dataset* inicial
- Biblioteca *requests* para requisições na api de notícias.
- Api de notícias utilizada: *newsapi*
- Biblioteca *deep_translator* para tradução das notícias
- Biblioteca *textblob* para análise de sentimentos
- Biblioteca *pandas_datareader* para buscar os preços das ações
- Informações das ações, *yahoo finance*

Para executar o projeto, entre na pasta *src* e depois execute a instrução *python3 main.py*

4 Desenvolvimento

O algoritmo desenvolvido para a análise de sentimentos de notícias financeiras faz um treinamento com um *dataset* de notícias do mercado. Esse treinamento é realizado pelo modelo de classificação de *Naive Bayes* usando a biblioteca *TextBlob*. Após o treinamento, ocorre uma busca de notícias financeiras em um intervalo de uma semana. As notícias são referentes à bolsa de valores oficial do Brasil, Bovespa. Além das notícias buscadas por meio da api, os valores referentes ao índice Bovespa também são buscados, porém esse por meio dos dados fornecidos pelo *yahoo finance*.

Todas as notícias passam por um processamento via *TextBlob* e, partir desse processamento, uma nota probabilística é obtida, essa nota varia entre 0.0 e 1.0. São avaliadas 20 notícias e 24 valores para o índice Bovespa. Com esses dados em mãos, é possível obter as seguinte informações:

Observando a figura 4, consegue-se perceber que durante estes 24 dias o índice Bovespa teve uma volatilidade acima do normal, no entanto, o objetivo deste trabalho é analisar a correlação entre os dados obtidos entre as figuras 5 e 4 e não a volatilidade da bolsa. Quanto à figura 5, o resultado

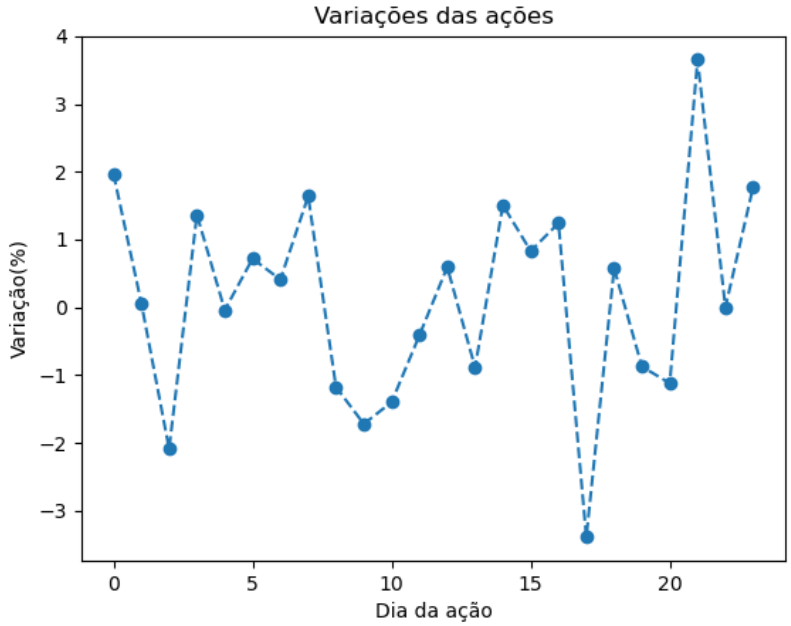

Figure 4: Variação do índice Bovespa para 24 dias

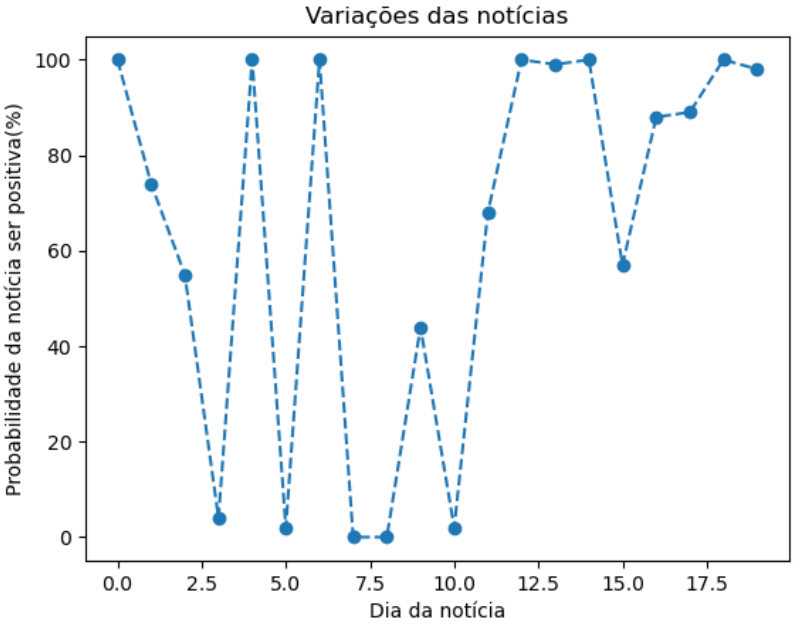

Figure 5: Notas obtidas a partir da análise de sentimentos para 20 notícias financeiras

obtido na análise de sentimentos de cada notícia teve resultados polarizados, ou seja, em diversas ocasiões a notícia foi avaliada como extremamente positiva, com o valor 100 ou próximo de 100.

Porém em alguns casos obteve-se notícias classificadas como neutra, como ocorre nos dias 2, 9 e 15. Em 5 ocasiões obteve-se notícias classificadas como negativas, ou seja, valores abaixo de 40%.

Um tipo de análise financeira que é muito usado por casas de investimentos é a correlação de ativos. Ela é uma medida estatística que mede a ligação entre duas variáveis, para o caso deste trabalho, as variáveis são os valores do índice Bovespa e as notícias relacionadas ao índice. A correlação é um valor número que varia entre 1.0 e -1.0. Quanto mais próximo do valor 1.0, maior é a relação entre essas duas variáveis, ou seja, se o índice Bovespa subir, maior a quantidade de notícias positivas relacionadas à ela. Agora, quanto mais próximo do valor -1.0, menor será a relação entre as duas variáveis, por exemplo, se o índice Bovespa subir, a quantidade de notícias negativas serão maiores. A biblioteca *numpy* possui abstrações que faz o cálculo da correlação entre duas variáveis, aplicando estes conceitos com base nas figuras 5 e 4, é possível obter o gráfico na figura 6.

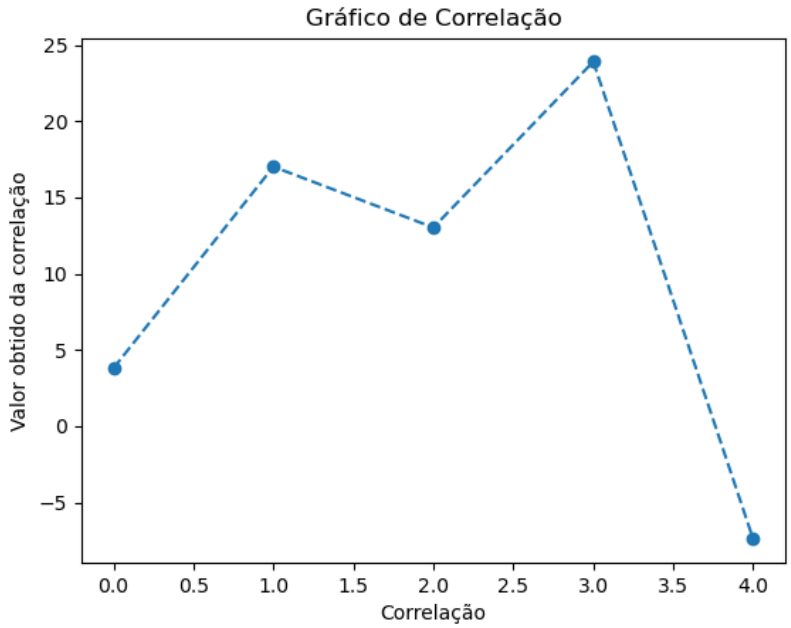

Figure 6: Correlação obtida a partir dos resultados obtidos para as notícias e índice Bovespa (%)

A figura 6 apresenta cinco pontos de correlação entre os ativos. Ao que tudo indica, as notícias possuem uma certa relação com o índice, porém é uma relação considerada baixa. O ponto que mais obteve correlação foi no terceiro ponto, onde a correlação foi perto da casa dos 25%. Em contrapartida, no quarto ponto, a correlação foi um valor negativo, próximo de -10%. Ou seja, as notícias mesmo que em baixa escala, andaram em sentido contrário ao índice no quarto ponto.

5 Conclusões

Prever o mercado não é uma tarefa fácil e muito investidores acreditam que isso é uma tarefa impossível. Porém, apesar das opiniões alheias, estudar o comportamento do mercado financeiro é uma

importante escada para atingir resultados positivos quando o assunto é resultado financeiro. No trabalho proposto, percebeu-se que existe uma correlação entre ativos e notícias financeiras. Portanto, é possível dizer que notícias financeiras podem, e provavelmente, alteram o humor dos investidores. Considerando que existem milhões de informações disponíveis na internet, o trabalho implementado foi apenas uma demonstração do uso de ferramentas de inteligência artificial e estatística. Portanto, considere que este trabalho serve apenas como uma demonstração de ferramentas da computação e não como um estudo profundo na área de finanças.

Um meio para obter resultados mais expressivos é utilizando outras formas de adquirir dados, treiná-los por diversas outras formas de implementação e o mais importante, estudar o histórico completo do índice, e não apenas um pequeno pedaço do calendário.

References

R. Gandhi. Naive bayes classifier. `https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c`. Accessed: 2021-12-07.

A. Mill. *Tudo o que você precisa saber sobre economia*. 3. Gente, 2017. ISBN 8545201699.

S. J. Russel and P. Norvig. *Inteligência Artificial*. 3. Campus, 2010. ISBN 978-0136042594.