À Coordenação de Trabalho de Conclusão de Curso de Informática
Departamento de Informática
Centro de Tecnologia
Universidade Estadual de Maringá

Encaminhamos, em anexo, a monografia intitulada "Análise de sentimento: Um estudo da correlação de notícias e os valores das ações de empresas" para avaliação desta coordenação.

Aluno: Eduardo H. K. Shibukawa

Orientador: Prof. Dr. Wagner Igarashi

Maringá, 20 de Novembro de 2017.

Universidade Estadual de Maringá
Centro de Tecnologia
Departamento de Informática

Análise de sentimento: Um estudo da correlação de notícias e os valores das ações de empresas

Eduardo H. K. Shibukawa

TCC-XX-07

Maringá, 20 de Novembro de 2017.

Universidade Estadual de Maringá
Centro de Tecnologia
Departamento de Informática

Análise de sentimento: Um estudo da correlação de notícias e os valores das ações de empresas

Eduardo H. K. Shibukawa

TCC-XX-07

Trabalho de Conclusão de Curso apresentado ao Curso de Informática, do Centro de Tecnologia, da Universidade Estadual de Maringá.
Orientador: Prof. Dr. Wagner Igarashi

Maringá, 20 de Novembro de 2017.

Eduardo H. K. Shibukawa

Análise de sentimento: Um estudo da correlação de notícias e os valores das ações de empresas

Este Trabalho de Conclusão de Curso foi julgado aprovado como requisito parcial para a obtenção do Grau de Bacharel em Informática, pela Universidade Estadual de Maringá.

Maringá, 20 de Novembro de 2017.

Banca Examinadora:

Dr. Wagner Igarashi
Universidade Estadual de Maringá (UEM)

Dr. Yandre Maldonado e Gomes da Costa
Universidade Estadual de Maringá (UEM)

Dra. Valéria Delisandra Feltrim
Universidade Estadual de Maringá (UEM)

Agradeço pelo mundo estar em constante transformação,
pelas coisas nunca serem da mesma forma,
pois assim não teríamos o que pesquisar,
o que descobrir e o que fazer.

Agradecimentos

Aos meus pais Jaime e Elsa, que sempre me apoiaram e deram forças, para terminar este curso. As amizades conquistadas durante essa vivência nesses anos da faculdade em que compartilhamos conhecimentos e experiências. A todos meus amigos e familiares que me apoiaram mesmo durante minha ausência. A todos as pessoas que diretamente ou indiretamente colaboraram para realização desse trabalho em especial ao meu orientador Prof. Dr. Wagner Igarashi, que me ajudou me orientado semanalmente durante este trabalho.

Live as if you were to die tomorrow.
Learn as if you were to live forever.

Mahatma Gandhi

Resumo

O mercado de ações é conhecido por ser um investimento de alto risco com uma grande volatilidade nos valores de seus títulos, com o propósito de prever os valores dessas ações existem duas escolas de pensamento que estudam o mercado financeiro, a escola de análise técnica e a escola de análise fundamentalista, simplificadamente a primeira escola tem como abordagem o estudo á partir de gráficos, buscando encontrar padrões para a compra e venda de ações, mais utilizada para investimentos de curto prazo, enquanto a segunda escola tem como abordagem determinar o valor das ações a partir de fatores externos que afetam o negócio de empresa e suas perspectivas futuras, assim como demonstrações financeiras da empresa para saber sua situação atual no mercado. Dentro deste contexto o trabalho atual tem como foco um dos itens da análise fundamentalista, onde temos como objetivo descobrir se existe uma possível correlação entre os sentimentos de notícias e seu impacto nos valores das ações, assim como a possível classificação de sentimentos de notícias utilizando técnicas da inteligência artificial. Para atingir tal objetivo foi implementado protótipos para cada etapa necessária sendo eles: a importação de valores de ações, a extração de notícias, a rotulação das notícias e a análise da correlação dos sentimentos com os valores da ações. Onde utilizamos dados referentes a empresa Petrobras, no período de maio a outubro de 2017. A análise desses dados mostrou uma correlação baixa, porém consistente entre os sentimentos das notícias com os valores das ações da empresa e período referentes. Com os resultados conclui-se que maís parâmetros devem ser levados em consideração na correlação, tais como peso maior em algumas notícias, um conjunto maior de dados abrangendo mais empresas, com um período maior nos dados extraídos.

Palavras-chave: Análise fundamentalista, Classificação Sentimento, Correlação Sentimento vs Valor da ação

Abstract

The stock market is known to be a high risk investment with great volatility in you stock values, with the purpose of predicting the values of the stocks, there are two schools of thought that study the financial market, the technical analysis school and fundamental analysis school, simplifying the first school has as an approach that study charts, seeking to find patterns for buying and selling stocks, commonly used to short term investments, while the second school has as its approach to determine the value of the stocks from external factories that that affect the company business and its future prospects, as well as financial demonstrations of the company to know its current situation in the market. Within the context the current work focuses on one of the items of fundamental analysis, where ou goal is to find out whether there is a possible correlation between news feelings and their impact on stock values, as well as classifying sentiments news using artificial intelligence techniques. To achieve such an objective there were implemented prototypes for each stage of the project: an importer of stock values, a news extractor, a news labeler, and an analysis of the correlation of feelings with stock values. Where utilized data from the Petrobras company, from May to October 2017. An analysis of those data showed a low but consistent correlation between the news sentiments and company stock values in this period. The results lead us to conclud that more parameteres need to be taken into account on the correlation, such as greater weight in certains news, a larger set of data covering more companies, a longer period of extracted data.

Palavras-chave: Fundamental analysis, Sentiment classification, Correlation Sentiment vs Stock Value

Lista de Figuras

FIGURA 1 – Algumas definições de inteligência artificial, organizadas em quatro categorias. (RUSSELL; NORVIG, 2013) 17
FIGURA 2 – Um modelo geral de agente com aprendizagem (RUSSELL; NORVIG, 2013) 18
FIGURA 3 – (a) Probabilidades a posteriori $P(h_i|d_1, \ldots, d_n)$, o número de observações N varia de 1 a 10, e cada observação é de um doce de lima. (b) Previsão bayesiana, $P(d_{n+1} = lima|d_1, \ldots, d_n)$, (RUSSELL; NORVIG, 2013) 23
FIGURA 4 – Arquitetura scrapy (SCRAPY, 2017) 26
FIGURA 5 – Arquitetura do projeto 29
FIGURA 6 – Programa de rotulação dos dados 31
FIGURA 7 – Arquivo bovespa 34
FIGURA 8 – Arquivo JSON extraido 35
FIGURA 9 – Matriz de confusão das métricas 37
FIGURA 10 –Gráficos de Ação X Sentimento - Normalizado 39
FIGURA 11 –Gráficos de mapas de calor dos campos de correlação e p-value . . . 40

Lista de Tabelas

Tabela 1 – Registro SQL 34
Tabela 2 – Arquivo CSV extraido 35
Tabela 3 – Dados rotulados 36
Tabela 4 – Sentimento lexicon 36
Tabela 5 – Sentimento scikit 37
Tabela 6 – Métricas do modelo 38
Tabela 7 – Correlação de sentimento vs valor da ação - Normalizado 39

Sumário

1 Introdução **12**

1.1 Justificativa 12

1.2 Motivação 13

1.3 Objetivos 13

1.3.1 Objetivo Geral: 13

1.3.2 Objetivo Específicos: 13

1.3.3 Organização do trabalho 13

2 Fundamentação Teórica **15**

2.1 Mercado Financeiro 15

2.2 Mercado De Ações 15

2.2.1 Análise técnica e fundamentalista 16

2.3 Inteligência Artificial 17

2.4 Processamento de Linguagem Natural 18

2.5 Análise de sentimento 19

2.6 Aprendizado de máquina 21

2.7 Aprendizagem Bayesiana 22

2.8 Trabalhos Correlatos 24

3 Desenvolvimento **25**

3.1 Materiais 25

3.1.1 Python 25

3.1.2 Scrapy 25

3.1.3 Scrapy Cloud 27

3.1.4 NLTK 27

3.1.5 Scikit-learn 28

3.1.6 TextBlob 28

3.1.7 Scipy 28

3.2 Projeto 28

3.2.1 Importação dos dados de ações 29

3.2.2 Extração de notícias 30

3.2.3 Rotulação dos dados 31

3.2.4 Análise de sentimento das notícias 32

3.2.5 Correlação dos dados 33

4 Resultados **34**

4.1 Importação dos dados de ações 34

4.2 Extração de notícias 35

4.3 Rotulação dos dados 36

4.4 Análise de sentimento das noticias 36

4.5 Correlação dos sentimentos com as ações 38

5 Desafios e Dificuldades **42**

6 Conclusão e Trabalhos Futuros **43**

6.1 Conclusão 43

6.2 Trabalhos Futuros 44

1 Introdução

1.1 Justificativa

Em meio à crise que nossa economia vem passando, muitas pessoas têm procurado outros meios de conseguir renda extra, uma dessas formas são os investimentos, Segundo Fonseca (2009), "Um projeto de investimento pode ser definido como um conjunto de informações que, quando reunidas, possibilitam uma tomada de decisão. Essa tomada de decisão está relacionada à alocação de recursos", existem vários tipos de investimentos, porém falaremos especificamente sobre as ações.

As ações são um tipo de investimento no qual se necessita de conhecimento financeiro e, além disso, do conhecimento sobre a influência do mercado nas empresas em que se quer investir. Na questão da influência do mercado nas empresas, este trabalho irá focar o impacto das notícias na variação de preço de ações.

Neste aspecto nós utilizaremos algumas subáreas da inteligência artificial o processamento de linguagem natural em conjunto com a Análise de Sentimento, também chamada de mineração de opinião, que é uma área de estudo que tem como objetivo classificar opiniões, sentimentos, avaliações, atitudes ou emoções em relação a produtos, serviços, organizações, problemas, etc, (LIU, 2012). Outra técnica que pode nos auxiliar neste processo é a aprendizagem de máquina, que segundo Monard e Baranauskas (2003) tem como objetivo a construção de sistemas capazes de adquirir conhecimento de forma automática. Os algoritmos de aprendizagem de máquina que buscam simular o raciocínio indutivo, melhorando seu desempenho na solução de certos problemas a partir da obtenção de exemplos.

Para a execução da classificação de sentimento precisa-se antes fazer a extração dos dados, que no contexto do trabalho serão notícias relacionadas às empresas selecionadas. Normalmente se é utilizado um *Web Crawler*, um robô que faz a extração automática de dados da web.

Com os resultados da classificação do sentimento de notícias de empresas de um período de tempo pré-definido, bem como da obtenção da variação dos preços de ações das respectivas empresas, será possível delinear o nível de correlação entre as variáveis envolvidas.

1.2 Motivação

Vivemos na era da informação, onde a cada segundo que passa uma enorme quantidade de dados é gerada, segundo a IBM (2017), "Todo dia, são gerados 2.5 Quintilhão de bytes de dados, é uma quantidade tão grande que nos 90% dos dados no mundo foram gerados nos últimos 2 anos". Porém para esses dados terem algum valor para as pessoas e empresas ele precisa ser analisado, uma das formas para a análise de dados é o uso da inteligência artificial.

Por motivos pessoais do acadêmico uma das áreas para a análise dos dados deste contexto seria o da bolsa de valores, e como elas podem ser diretamente afetadas por notícias no dia a dia. Assim justificando-se um estudo, utilizando-se da análise de sentimento das notícias e sua correlação com os valores das ações na bolsa de valores.

1.3 Objetivos

1.3.1 Objetivo Geral:

Análise de sentimentos de notícias, com finalidade de verificar a possível correlação entre as notícias sobre uma empresa e as variações do preço de suas ações.

1.3.2 Objetivo Específicos:

Com base em nosso objetivo geral estabelecemos os seguintes objetivos específicos:

- Extração de notícias de sites.
- Análise de sentimento das notícias extraídas.
- Análise da correlação dos dados da análise de sentimento com os dados de valorização ou desvalorização das mesmas.

1.3.3 Organização do trabalho

A partir desta Seção, que tem como caráter introdutório, o documento está organizado da seguinte maneira: A seção 2 de fundamentação teórica tem como propósito o embasamento teórico de nosso projeto, apresentando conceitos como o do mercado financeiro (2.1), sobre o mercado de ações (2.1) e as análises técnicas e fundamentalista (2.2.1), a inteligência artificial

(2.3), o processamento de linguagem natural (2.4), a análise de sentimento (2.5), o aprendizado de máquina (2.6), a aprendizagem Bayesiana (2.7) e para finalizar esta seção os trabalhos correlatos (2.8). A seção 3 apresenta o desenvolvimento do projeto, primeiro falamos sobre as ferramentas utilizadas (3.1) e após isso sobre o projeto (3.2). A seção 4 apresenta os resultados obtidos. A seção 5 apresenta os desafios e dificuldades encontrados no decorrer do projeto. Por ultimo a seção 6 apresenta a conclusão do trabalho (6.1) e os trabalhos futuros (6.2).

2 Fundamentação Teórica

2.1 Mercado Financeiro

Antes de falar sobre o mercado financeiro vamos definir o que é a economia, segundo o dicionario, "Ciência que trata da produção, distribuição e consumo das riquezas de uma nação."(AULERIO, 2017), além disso Fischer et al. (2014) estrutura a economia da seguinte forma, os agentes econômicos, são famílias, empresas ou governo que compõe o sistema econômico moderno. As famílias oferecem insumos necessários pela empresa, em troca de rendimentos que são salários, juros, lucros e alugueis. Com isso as famílias conseguem comprar os produtos ou serviços oferecidos pelas empresas, e o governo recolhe impostos das famílias e empresas, e devolve para a sociedade em forma de projetos ou serviços sociais.

Segundo Selan (2015) o sistema financeiro é composto por instituições econômicas que tem como objetivo a intermediação entre poupadores e investidores. Ainda, segundo Bertolo (2002), os poupadores depositam e emprestam dinheiro para instituições econômicas, e estas utilizam este dinheiro para financiar alguns setores da economia que estão precisando de recursos.

O mercado de capitais faz parte do mercado financeiro e, neste, os poupadores destinam seus recursos diretamente ao desenvolvimento econômico de forma direta. Por exemplo, empresas que precisam de recursos conseguem financiamento, por meia da emissão de títulos vendidos diretamente aos poupadores, que agora podem ser chamados de investidores. O mercado de capitais pode ser divido entre cinco tipos, sendo eles: Mercado de renda variável; Mercado de renda fixa; Mercado de câmbio; Mercado de derivativos; Mercado de fundos de investimento. Em nosso trabalho iremos focar na análise de ações que faz parte do mercado de renda variável (SELAN, 2015).

2.2 Mercado De Ações

Segundo Bertolo (2002), as ações são a menor parcela de capital de uma empresa. E elas são títulos que não garantem o retorno dos recursos alocados pelos investidores, uma vez que sua remuneração é determinada pela capacidade da empresa em gerar lucros.

Ao se comprar ações, os investidores se tornam sócios da empresa, e caso o investidor mudar de opinião, quanto a capacidade da geração de lucro da empresa, ele pode vende-las.

Essa negociação é feita pelo intermédio da bolsa de valores, e seu funcionamento é regido pela comissão de valores mobiliários (FISCHER et al., 2014)

As ações podem ser divididas em dois tipos, as ordinárias e as preferenciais. Na primeira o investidor tem o direito de voto em assembleias de acionistas da empresa, na segunda ele não possui esse direito (BERTOLO, 2002).

O preço das ações está relacionado diretamente à sua oferta e procura. Quanto maior sua a procura mais seu preço irá aumentar. Sua procura está relacionada a vários fatores tais como: político, econômico, estratégias das empresas, inovações e aumento de competitividade da empresa no mercado, etc. (BERTOLO, 2002), para a compra e venda destas ações é comum se utilizar a análise técnica, ou a fundamentalista.

2.2.1 Análise técnica e fundamentalista

Segundo Murphy (1999), a análise técnica se concentra no estudo do mercado de ações, enquanto na análise fundamentalista o foco são fatores externos, indicadores financeiros da empresa, o mercado econômico, etc, que fazem a demanda pelas ações aumentarem ou diminuírem, assim causando alterações no preço das ações.

Na abordagem fundamentalista, existe uma grande diferença entre o valor das ações da empresa, e seu verdadeiro ou potencial valor, também chamado de valor intrínseco, ele é calculado a partir de vários fatores relevantes do mercado econômico, fatores operacionais da própria empresa, tais como receita, custos, etc, e tentam avaliar como tais fatores pesam no valor real da empresa.

Enquanto na abordagem técnica, se utiliza de varias ferramentas, tais como indicadores e desenhos gráficos na identificação de tendências. As técnicas gráficas tem como objetivo encontrar um padrão de crescimento do histórico de preços, a fim de identificar um bom momento para compra e venda das ações.

Neste trabalho, tem-se como foco a análise fundamentalista, utilizando as notícias como um fator, e verificando se elas possuem alguma correlação com o preço das ações, com o uso de técnicas da inteligência artificial, como a análise de sentimento, e o aprendizado de máquina.

2.3 Inteligência Artificial

Segundo Russell e Norvig (2013), a inteligência artificial pode ser dívida em quatro estratégias diferentes, as estratégias e suas definições podem ser vistas na figura 1.

Pensando como um humano	**Pensando racionalmente**
"O novo e interessante esforço para fazer os computadores pensarem (...) *máquinas com mentes*, no sentido total e literal." (Haugeland, 1985) "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985) "O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992)
Agindo como seres humanos	**Agindo racionalmente**
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzweil, 1990) "O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas." (Rich and Knight, 1991)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes." (Poole *et al.*, 1998) "AI... está relacionada a um desempenho inteligente de artefatos." (Nilsson, 1998)

Figura 1: Algumas definições de inteligência artificial, organizadas em quatro categorias. (RUSSELL; NORVIG, 2013)

Neste trabalho, tem-se como foco a última abordagem "Agindo Racionalmente", ela tem como base o uso de agentes inteligentes.

> "Um agente é simplesmente algo que age (a palavra agente vem do latino agere, que significa fazer). Certamente todos os programas de computador realizam alguma coisa, mas espera-se que um agente computacional faça mais: opere sob controle autônomo, perceba seu ambiente, persista por um período de tempo prolongado, adapte-se a mudanças e seja capaz de criar e perseguir metas. Um agente racional é aquele que age para alcançar o melhor resultado ou, quando há incerteza, o melhor resultado esperado.", (RUSSELL; NORVIG, 2013)

Segundo Poole e Mackworth (2010), a IA é a área que estuda a síntese e a análise de agentes computacionais que agem inteligentemente, agentes são inteligentes quando executam ações apropriadas para as circunstâncias e seus objetivos, são flexíveis para mudanças de ambientes e mudanças de objetivos, aprendem com a experiência.

Existem vários tipos de agentes; Agentes reativos simples; Agentes reativos em; modelos; Agentes baseados em objetivos; Agentes baseados na utilidade; Agentes com aprendizagem

Em nosso trabalho utilizaremos a ideia de agentes com aprendizagem, ele pode ser divido em quatro componentes conceituais, que podem ser vistos na figura 2.

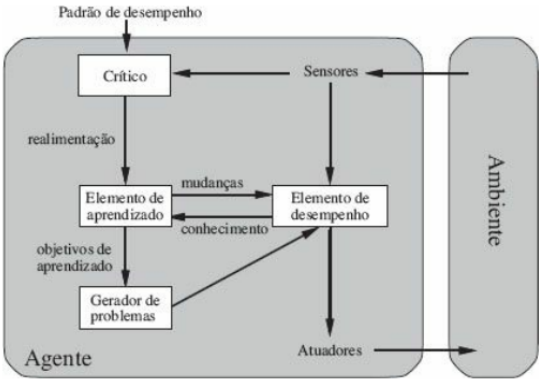

Figura 2: Um modelo geral de agente com aprendizagem (RUSSELL; NORVIG, 2013)

Neste modelo os dois principais componentes são o elemento de aprendizado que é responsável pela execução de aperfeiçoamentos, e o elemento de desempenho, que recebe percepções e decide sobre as ações que se vai tomar. Utilizaremos esse conceito de agentes de aprendizagem para a resolução de nosso problema de processamento da linguagem natural.

2.4 Processamento de Linguagem Natural

Antes de definir o processamento natural, devemos definir o que é a linguagem, de acordo com o dicionário Aurélio temos como definição de linguagem

> "Linguagem formal, linguagem simbólica que serve de axiomas e leis, bem como de normas especiais, em opos. à linguagem natural; Linguagem natural, o conjunto de sinais que se empregam e interpretam indistintivamente (como a fala, o grito, os olhares, os gestos etc.); Faculdade que têm os homens de comunicar-se uns com os outros, exprimindo seus pensamentos e sentimentos por meio de vocábulos, que se transcrevem quando necessário Voz, grito, canto dos animais: linguagem dos papagaios.", (AULERIO, 2017)

Segundo Russell e Norvig (2013), o processamento de linguagem natural tem como objetivo adquirir conhecimento ao menos parcial, sobre a linguagem que os humanos usam, o problema é que a linguagem natural é ambígua e complexa e está em constante mutação.

Segundo Jurafsky e Martin (2009), para a resolução deste complexo problema devemos separar/analisar a linguagem natural em seis categorias: Fonética e fonologia: O estudo da linguagem da fala; Morfologia: O estudo dos componentes significativos das palavras; Sintaxe: O estudo das estrutura e relações entre as palavras; Semântica: O estudo do significado das

palavras em seu contexto; Pragmática: O estudo de como a linguagem atinge seus objetivos; Discurso: O estudo de unidades linguísticas maiores que uma única expressão;

2.5 Análise de sentimento

Segundo Liu (2012), análise de sentimento, também chamado de mineração de opinião, é a área que estuda e analisa opiniões, sentimentos, avaliações, atitudes e emoções em relação a entidades como produtos, serviços, organizações, indivíduos, problemas, etc.

Para Tsytsarau e Palpanas (2012), a análise de sentimento é uma análise de subjetividade mais refinada, as duas possuem a mesma essência, a análise de subjetividade tem como objetivo classificar conteúdos em objetivo ou subjetivo, e é composta de três passos, a identificação, classificação e a agregação, já a análise de sentimento tem como função identificar a opinião expressada sobre um assunto em particular, e avaliar a polaridade desta expressão, no caso distinguir se está é positiva ou negativa.

Segundo Liu (2012), as pesquisas relacionadas a análise de sentimento, são separadas por níveis de acordo com sua granularidade, seus principais níveis são:

- Nível de documento: tem como objetivo classificar a opinião de um documento inteiro, este nível de análise leva em consideração que cada documento expressa uma opinião sobre uma única entidade.
- Nível de sentença: tem como objetivo analisar se a sentença foi expressada como positiva, negativa ou neutra, sendo fortemente relacionada com a análise de subjetividade citada anteriormente.
- Nível de entidade ou aspecto: é baseada na ideia de que uma opinião consiste em um sentimento positivo ou negativo, e um alvo. Seu objetivo é descobrir o sentimento de entidades e suas diferentes características, é o nível mais complexo de análise.

Para a resolução de nosso problema precisamos definir o que é uma opinião em nosso contexto. A opinião pode ser classificada em dois tipos, a opinião regular e a opinião de comparação. No primeiro tipo de opinião é expressado um sentimento sobre um alvo, já no segundo tipo é expressando um sentimento de um alvo em relação a outro alvo. Na literatura a opinião regular é citada regularmente como opinião, em nosso trabalho iremos nos focar nesse tipo de opinião.

A opinião consiste em dois componentes principais, um alvo e o sentimento em relação ao alvo ou (g, s) , onde g é a entidade sobre a qual alguma opinião foi expressada, e s é um sentimento positivo, negativo ou neutro, podendo ser também uma avaliação numérica expressando a intensidade do sentimento.

Hu e Liu (2014) definem a opinião como uma tupla de cinco elementos $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, onde e_i é o nome da entidade, a_{ij} é um aspecto de e_i, s_{ijkl} é o sentimento do aspecto a_{ij} da entidade e_i, h_k é o detentor da opinião, e t_l é o tempo quando a opinião foi expressada por h_k.

Com essa definição conseguimos definir os objetivos e as principais tarefas da análise de sentimento. Nosso objetivo pode ser definido como: dado um documento de opinião d queremos descobrir todas opiniões em d, onde cada opinião pode ser representada pela tupa $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$.

Para facilitar nossa explicação iremos definir modelos de entidade e e documento de opinião d:

- Modelo Entidade: Uma entidade e_i é representada como um conjunto finito de aspectos $A_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$. Podendo ser representada como qualquer uma das entidades dentro do conjunto final de entidades de expressões $OE_i = \{o_{ei1}, o_{ei2}, \ldots, o_{ein}\}$. Cada aspecto $a_{ij} \in A_i$ da entidade pode ser representada como aspecto de expressão contido no conjunto finito de aspectos de expressões $AE_{ij} = \{a_{eij1}, a_{eij2}, \ldots, a_{eijm}\}$
- Modelo Documento de opinião: Um documento de opinião d contém um conjunto de entidades $\{e_1, e_2, \ldots, e_r\}$, e um subconjunto de seus aspectos de um conjunto de detentores de opiniões, em um determinado momento no tempo.

Resumidamente, dado um conjunto de documentos D, a análise de sentimento, consiste em 6 tarefas principais.

A Primeira tarefa é a extração de entidade e categorização, que consiste na extração de todas entidades de expressão, e categorização ou agrupamento de entidades de expressões sinônimas, em suas respectivas categorias. Cada categoria representa uma única entidade e_i.

A segunda tarefa é a extração de aspectos e categorização, que consiste na extração de todos os aspectos de expressão das entidades e a categorização de seus aspectos de expressão em seus *clusters*. Cada *cluster* do aspecto de expressão da entidade e_i representa um aspecto a_{ij}.

A terceira tarefa é a extração do detentor de opinião e categorização, que consiste na extração dos detentores de opinião do texto ou dos dados estruturados, e sua categorização.

A quarta tarefa é a extração de tempo e uniformização, que consiste na extração do tempo na qual opiniões foram obtidas, e na uniformização desses dados.

A quinta tarefa é a classificação de sentimento do aspecto, que consiste em determinar se a opinião em um aspecto a_{ij}, é positivo, negativo ou neutro, ou atribuir sua avaliação numérica ao sentimento do aspecto.

A sexta e última tarefa é a geração da tupla de cinco elementos, que consiste em produzir todas tuplas de cinco elementos $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ contidas no documento d baseado no resultado das tarefas acima.

O modelo apresentado acima, pode ser considerado um *framework* para obtermos os dados necessários para a análise do sentimento dos documentos e dos aspectos alvos. Para a tarefa 5, classificação dos sentimentos do aspecto é necessário o uso de alguma ferramenta para ajudar a resolver o problema de classificação, em nosso caso iremos nos focar no aprendizado de máquina, que tem esse intuito como um dos seus objetivos.

2.6 Aprendizado de máquina

De acordo com Barber (2016):

> "Machine learning is the body of research related to automated large-scale data analysis. Historically, the field was centred around biologically inspired models and the long term goals of much of the community are oriented to producing models and algorithms that can process information as well as biological systems. The field also encompasses many of the traditional areas of statistics with, however, a strong focus on mathematical models and also prediction. Machine learning is now central to many areas of interest in computer science and related large-scale information processing domains.", (BARBER, 2016)

Segundo ele o aprendizado de máquina é fundamentalmente sobre extrair dados de grandes conjuntos de dados, muitas vezes tem como motivação produzir algoritmos capazes de imitar ou melhorar a performance humana.

A área de aprendizagem de máquina pode ser dívida em várias subáreas, os dois principais campos da área podem ser considerados o de aprendizado supervisionado e o aprendizado não supervisionado. A área de aprendizado supervisionado tem o foco de melhorar a acurácia dos

resultados da predição, enquanto o outro encontrar de descrições plausíveis sobre os dados, iremos manter o foco no aprendizado supervisionado.

Nosso foco neste trabalho será no aprendizado supervisionado, que segundo Poole e Mackworth (2010), pode ser abstraído como mostrado a seguir. Assumindo que seja disponibilizado os seguintes dados ao nosso classificador. Um conjunto de atributos de entrada, $x_1, x_2, \ldots, x_n$; Um conjunto de atributos alvos, $y_1, y_2, \ldots, y_n$; Um conjunto de dados de treino, onde os atributos de entrada e atributos alvos são fornecidos para cada exemplo; Um conjunto de dados de teste, onde apenas os atributos de entrada são fornecidos;O objetivo é prever os valores dos atributos alvos, com base nos exemplos de teste, e de dados ainda não fornecidos. Aprendizado é a criação de uma representação que consegue fazer predições baseado nos atributos de entrada de novos exemplos.

Outra visão da aprendizagem supervisionada é a de Russell e Norvig (2013), onde dado um conjunto de treinamento com n pares de entrada x e uma saída y, onde toda saída foi gerada por uma função desconhecida $y = f(x)$, e nosso objetivo é descobrir uma função hipótese h que se aproxime da função $f(x)$.

Para a escolha da melhor hipótese h, existe vários modelos de algoritmos, tais como o de regressão linear, as redes neurais, as máquinas de vetor de suporte, o modelo Bayesiano, entre outros, em nosso trabalho decidimos, escolher o modelo Bayesiano, para a resolução do problema de classificação do sentimento das notícias extraídas.

2.7 Aprendizagem Bayesiana

Segundo Poole e Mackworth (2010) a ideia do aprendizado bayesiano é computar posteriormente a distribuição de probabilidades, das características de novos exemplos, a partir destes novos exemplos em conjunto com todos os exemplos posteriores.

Para Russell e Norvig (2013), a aprendizagem bayesiana, simplesmente calcula a probabilidade de cada hipótese, ela faz previsões de acordo com os dados posteriores. Ao invés de utilizar apenas a melhor hipótese, as previsões são feitas com o uso de todas hipóteses, ponderadas por suas probabilidades.

Supondo que D é a representação dos exemplos posteriores, com valor observado d, a probabilidade de uma hipótese h_i pode ser obtida pela regra de Bayes, $P(h_i|d) = \alpha P(d|h_i)P(h_i)$, supondo que queremos fazer a previsão de um valor desconhecido X teremos,

$P(X|d) = \sum_i P(X|d, h_i)P(h_i|d) = \sum_i P(X|h_i)P(h_i|d)$, então pressupomos que cada hipótese determina uma distribuição de probabilidade sobre X. Com isso temos um equação que mostra que as previsões são médias ponderadas sobre as previsões das hipóteses individuais.

Para facilitar a visualização do funcionamento da aprendizagem bayesiana, vamos considerar um exemplo simples, onde temos um pacote de doces, que pode ter até dois sabores de doce dentro dele, cereja ou lima, os pacotes são sempre produzidos respeitando uma respectiva relação, sendo elas, 100% cereja, 75% cereja e 25% lima, 50% cereja e 50% lima, 25% cereja e 75% lima, 100% lima, vamos abreviar estas relações respectivamente de $\{h_1, h_2, h_3, h_4, h_5\}$, dado um novo pacote de doces, com uma hipótese aleatória H denotando o tipo do pacote de doces, com valores possíveis de h_1 até h_5., e a medida que os doces são retirados do pacote, são revelados os dados $\{D_1, D2, \ldots, D_n\}$, onde cada D_i é um valor aleatório que contem o sabor do doce, sendo ele cereja ou lima, temos como objetivo prever o sabor do próximo doce retirado do pacote, sabendo que a hipótese a priori sobre $\{h_1, h_2, h_3, h_4, h_5\}$, é {10%, 20%, 40%, 10%, 20%}.

Com o uso da formula de Bayes, supondo que os dados são independentes e identicamente distribuídos, conseguimos calcular a probabilidade de algum dado com a formula, $P(d|h_1) = \prod_j P(d_j|h_i)$, para facilitar o entendimento, levando em consideração que retiramos o doce do sabor de lima em todas nossas iterações, a figura figura 3(a) mostra como a probabilidade de ser cada uma das hipóteses muda a cada iteração em que os doces são retirados do pacote, podemos notar que como as probabilidades começam com seu valores a priori, a hipótese h_3 é a mais provável no inicio, e esse valor vai se alterando para a hipótese h_5 lentamente, já na figura 3(b), conseguimos observar a probabilidade do próximo doce ter do sabor de lima, como esperado a medida que os doces são retirados, a probabilidade dele ser de lima, vai aumentando constantemente em direção ao valor 1.

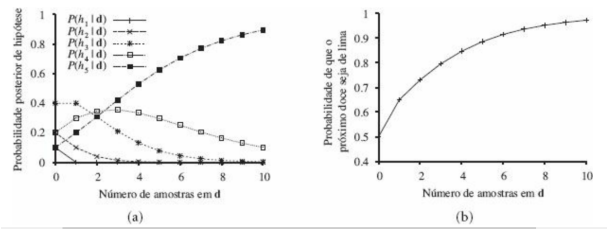

Figura 3: (a) Probabilidades a posteriori $P(h_i|d_1, \ldots, d_n)$, o número de observações N varia de 1 a 10, e cada observação é de um doce de lima. (b) Previsão bayesiana, $P(d_{n+1} = lima|d_1, \ldots, d_n)$, (RUSSELL; NORVIG, 2013)

2.8 Trabalhos Correlatos

Alguns trabalho estudados durante o processo de desenvolvimento de nossa monografia, possui grande correlações com a mesma, entre eles podemos citar três, sendo o primeiro a tese de doutorado de Alves (2015), Uso de técnicas de Computação Social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores, sua tese tem como objetivo, analisar a relação de dados obtidos da rede social twitter e o mercado de ações brasileiro, por meio de um sistema que auxilia a tomada de decisão de compra e venda. Para sua análise ela utiliza da escola técnicas, e compara eles se utilizando também da escola fundamentalista, onde utiliza o sentimento da base de dados do twitter.

O artigo de Mittal e Goel (2012), Stock prediction using twitter sentiment analysis, tem como objetivo, mostrar como os sentimentos do público do twitter influenciam a bolsa de valores, para isso eles utilizaram da análise de sentimento nos dados obtidos do twitter, com a ajuda de modelos de aprendizagem de maquina, eles conseguiram obter correlações entre os sentimentos de sua análise de sentimento e o preço das ações.

O artigo de Nuij et al. (2014), An automated framework for incorporating news into stock trading strategies, o autor tem como objetivo apresentar um *framework*, para a incorporação de notícias em estratégias de negociações de ações, eles agrupam as notíciass em eventos que influenciam a mudança dos valores das ações, e procuram observar uma correlação entre os eventos das notíciass extraídas e os preços das ações, após isso utilizam os eventos como fatores, em análises utilizando a escola técnica.

3 Desenvolvimento

Com o foco em nosso objetivo geral, dividimos nosso projeto em etapas, elas seriam o *scraping* ou extração de notícias, a importação e o armazenamento de dados sobre ações da bolsa de valores, o desenvolvimento de uma ferramenta para ajudar na rotulação dos dados extraídos, a análise de sentimento das notícias extraídas, a correlação do sentimento das notícias e o valores da bolsa, a plotagem dos dados obtidos e, a análise dos dados obtidos.

Ao longo da implementação destes protótipos, foram feitos diversos testes, e vários ajustes foram necessários, esta seção irá mostrar, todas atividades realizadas e ferramentas utilizadas para o desenvolvimento destes protótipos.

3.1 Materiais

3.1.1 Python

Python é uma linguagem de programação de alto nível e de código aberto, sendo gerenciada pela Python Software Foundation (PSF), é utilizada por muitas empresas para projetos reais, que estão em produção, em seu portfólio estão empresas de grande porte como até mesmo a Google. É muito utilizada para a análise de dados, pois existem *frameworks* prontos para o uso tais como o Scikit-learn, Spacy, NTLK, MatPlotLib, TensforFlow, etc.

3.1.2 Scrapy

Scrapy é um *framework* de código aberto, escrito na linguagem python, que tem como objetivo facilitar a criação de aplicações que realizam o *web crawling*, o *web crawling* é uma técnica que consiste em extrair informações de forma automática dos sites da internet, com a ajuda de robôs também chamados de *spiders*, para assim fazer a extração dos dados automaticamente, o que seria inviável de ser feito manualmente, devido a quantidade de dados necessária, para essa tarefa.

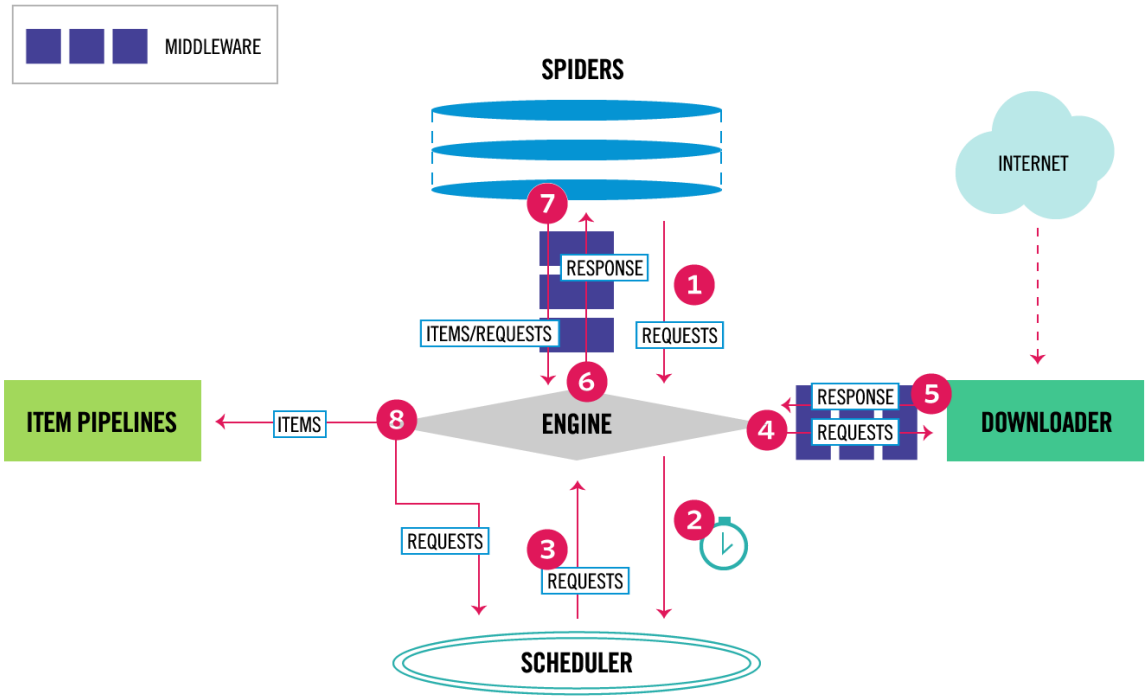

Figura 4: Arquitetura scrapy (SCRAPY, 2017)

A figura 4, ilustra de maneira geral a arquitetura do *framework*, ela é composta de vários componentes, sendo eles o motor, o agendador, o *downloader*, as *spiders*, o *pipeline* do item e, os *middlewares*, os componentes, seguido de seu fluxo de dados são explicado posteriormente.

O motor é responsável por controlar o fluxo de dados, entre todos os componentes do sistema, e ativar os gatilhos dos eventos quando certas ações ocorrerem; O agendador recebe requisições do motor e enfileira eles, para alimentá-los depois, quando o motor fizer suas requisições; O *downloader* é responsável por obter os dados das páginas web, e alimentar o motor com eles, que por sua vez irá alimentar as *spiders*. As *spiders* são classes customizáveis pelos usuários, que analisam uma resposta HTTP, e para então extrair seus itens, ou seguir outros respostas; O pipeline do item é responsavel por processar o item, após sua extração pelas *spiders*, é comumente feito tarefas como a limpeza dos dados, validação ou sua persistência em um banco de dados; Os *middlewares* podem ser de download, ou de *spider*, e são usados como uma interface de compatibilização da comunicação entre os componentes, para facilitarem a customização do requisição/resposta, ou no processamento de uma *spider*.

O controle do fluxo de dados segue o seguinte padrão, primeiro o motor busca a

requisição inicial para a extração dos dados na *spider*, então o motor agenda as requisições no agendador e requisita a próxima requisição para extrair, o agendador então retorna a próxima requisição para o motor, o motor envia a requisição para o *downloader*, passando pelos *middlewares* de download, assim que for finalizado o download da página, o *downloader* gera uma resposta e envia para o motor, passando novamente pelos middlewares de download, assim o motor recebe esta requisição e então envia para as *spiders* para o processamento, passando pelos *middlewares* da *spider*, a *spider* processa a resposta e retorna os itens extraídos, ou uma nova requisição para o motor e, passa novamente pelos *middlewares* das *spider*, o motor então envia os itens extraídos para o *pipeline* do item, que então envia os itens processados para o agendador, que requisita uma nova requisição, continuando este ciclo até acabar todas as requisições.

3.1.3 Scrapy Cloud

Scrapy Cloud é uma plataforma utilizada para a execução de *web crawlers*. Pode ser pensada como um Heroku para rastreamento na web.

O Heroku é uma plataforma como serviço (PasS) na nuvem, executada sobre a Amazon EC2 uma infraestrutura como serviço (IaaS), o Heroku utiliza um sistema operacional debian e possui a função de automatizar a criação de novas máquinas virtuais e configurar o ambiente para rodar sua aplicação, inicialmente suportava somente a linguagem de programação Ruby, porém atualmente suporta diversas linguagens de programação como Java, Node.JS, Scala, Clojure, Python e PHP.

3.1.4 NLTK

NLTK é um conjunto de bibliotecas escrito em python, que tem como objetivo o processamento de linguagem natural, foi desenvolvido por Steven Birde Edward Loper, no departamento de ciência da computação da universidade da Pensilvânia, é muito utilizada no meio acadêmico, ela suporta classificações, tokenização, *stemming*, *parseamento*, semântica e funcionalidades de raciocínio.

3.1.5 Scikit-learn

Scitkit-learn é uma biblioteca de código aberto, escrita na linguagem python, que auxilia no desenvolvimento de soluções para a análise de dados e a extração de dados. Construída em cima das bibliotecas NumPy, SciPy e MatPlotLib, contém vários algoritmos de aprendizagem de maquina já implementados para o uso.

3.1.6 TextBlob

TextBlob é uma biblioteca, escrita na linguagem python, que auxilia no processamento de dados textuais, fornecendo uma API simples, para tarefas comuns de processamento de linguagem natural, tais como análise de sentimento, classificação, tradução, etc.

3.1.7 Scipy

SciPy é uma biblioteca de código aberto, escrita na linguagem python, com a intenção de ser usadas por matemáticos, cientistas e engenheiros. Ela utiliza o NumPy como base, que fornece uma manipulação rápida e conveniente de um vetores N-dimensional, e implementa diversos algoritmos em C, para otimizar seu processamento, ela tem implementa funções estatísticas, de otimização, integração numérica, processamento de sinais e imagens, solução de equações diferenciais, funções especiais, e polinômios.

3.2 Projeto

Nesta seção iremos falar sobre as ferramentas utilizadas, sobre a arquitetura de nosso projeto, e o funcionamento de cada etapa.

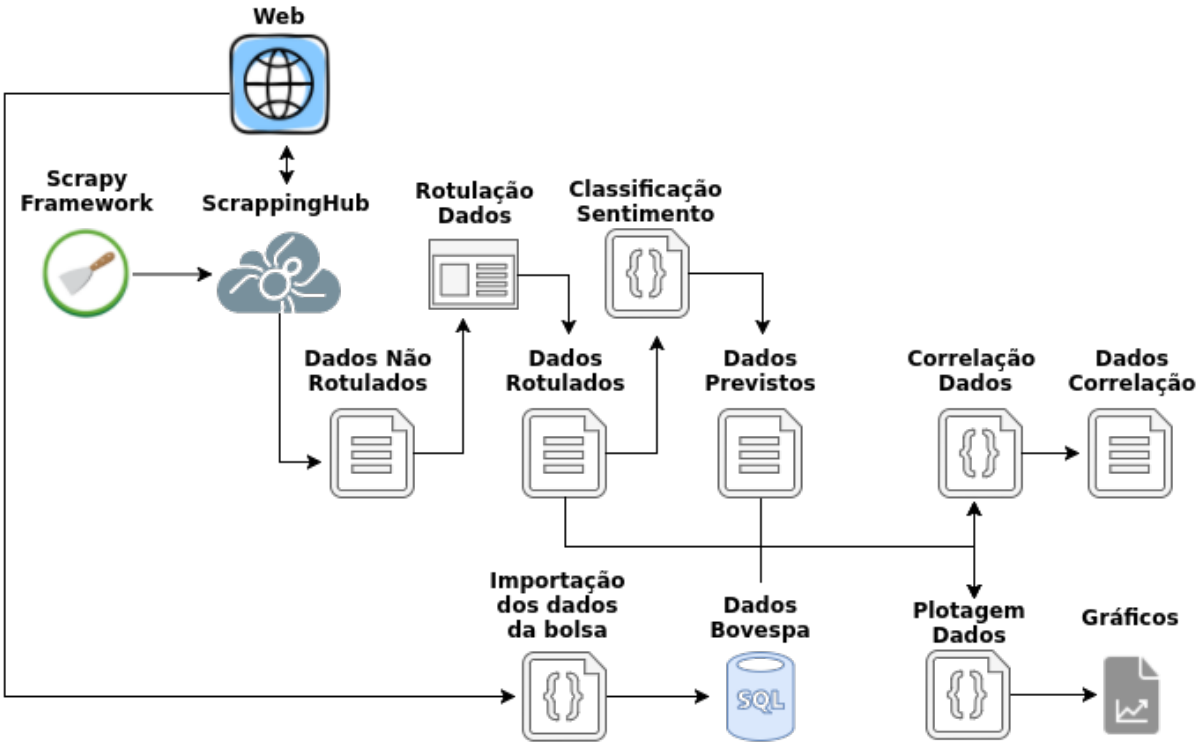

Figura 5: Arquitetura do projeto

A figura 5 mostrada anteriormente ilustra a arquitetura de nosso projeto, em nosso caso primeiramente temos a aplicação de importação de dados da bolsa que insere os dados em um banco de dados, após isso temos uma ferramenta criada utilizando o *framework* Scrapy em conjunto com a plataforma ScrappingHub, que tem a função de extrair as notícias não rotuladas. Em nossa próxima etapa temos um aplicativo que facilita a rotulação de dados que tem como entrada as notícias não rotuladas e como saída as notícias rotuladas. A seguir uma aplicação lê os dados rotulados e os classifica gerando dados com as previsões do sentimentos. Com isso nos resta calcular as correlações tanto dos dados rotulados quanto dos dados de previsões com os valores da bolsa que estão no banco de dados, gerando informações sobre a correlação dos dados, e gerando os gráficos dos dados para a facilitar nossa análise, todas essas etapas estão explicadas com mais detalhes nas seções posteriores.

3.2.1 Importação dos dados de ações

Para a importação dos dado das ações, utilizamos os arquivos providos pela entidade BM&FBOVESPA, em seu site conseguimos obter o histórico de ações em um arquivo com filtros diversificados, sendo eles por ano, por mês de um ano, ou por um dia específico, os dados desse arquivo podem ser interpretados seguindo um layout fornecido no próprio site, assim

seguindo esse layout, utilizando a biblioteca *peewee*, criamos um modelo estendendo a classe *peewee.Model* modelamos uma classe para as ações com o nome Registro001, que possui todos os campos especificados no layout, após isso fizemos a leitura do arquivo da BM&FBOVESPA, e mapeamos todos os campos do arquivo com o do nosso modelo, e filtramos somente as ações da Petrobras (PETR4) e armazenamos os dados em um banco de dados *SQLLite* nomeado bovespa.db, para ser usado posteriormente.

3.2.2 Extração de notícias

Para a realização de nosso protótipo precisamos de dados, assim utilizando a linguagem *python*, em conjunto com o *framework Scrapy*, criamos dois *crawlers* na plataforma *scrapinghub*, para extrair os dados dos sites de notícias G1, e EXAME. Seu desenvolvimento será descrito a seguir.

Para a obtenção das notícias, tivemos a ideia de utilizar a pesquisa do google, limitando o período de nossa pesquisa, e utilizando os parâmetros *allintitle:string* e *site:string*, conseguimos filtrar respectivamente, os termos pesquisados no título dos site e buscar as *URLs* somente de um domínio específico, assim conseguindo uma consulta com os links para a extração.

Para extrair os dados das notícias desta consulta, criamos um novo projeto utilizando o *framework Scrapy* e, seguido da criação de uma nova *spider* para este projeto, após isso normalmente é necessário implementar apenas a função *parse* para a extração dos dados, porém em nosso caso tivemos que implementar a função parse para extrair os *links* de toda consulta desde a pagina 1 até a N, após isso implementamos uma função *parse* para o domínio que queremos extrair os dados no caso a G1, em nosso *parser* buscamos basicamente 3 propriedades para nosso item, que são o título, a data de atualização, e o conteúdo da notícia, para isso utilizamos a linguagem de consulta *xPath*, para buscar os dados que queremos no HTML da requisição.

Para a execução de nosso projeto, fizemos um *deploy* na plataforma *Scrapy Cloud*, após isso executamos a extração de dados os parâmetros *site:*G1 e, *allintitle:* petrobras, com o período de 01/05/2017 á 31/08/2017, e outro com o com o período de 01/09/2017 á 30/09/2017, após isso acessamos os dados e baixamos no formato JSON e CSV respectivamente.

3.2.3 Rotulação dos dados

Para a análise de sentimento de notícias, utilizando a abordagem do aprendizado de máquina supervisionado, é necessário uma base de dados rotulados, esta seção irá explicar como foi o processo de rotulação.

Após extraídas as notícias, decidimos rotular os dados de maio a agosto, para isso decidimos criar um programa para facilitar a visualização dos dados, e a rotulação, ele pode ser visto na figura 6.

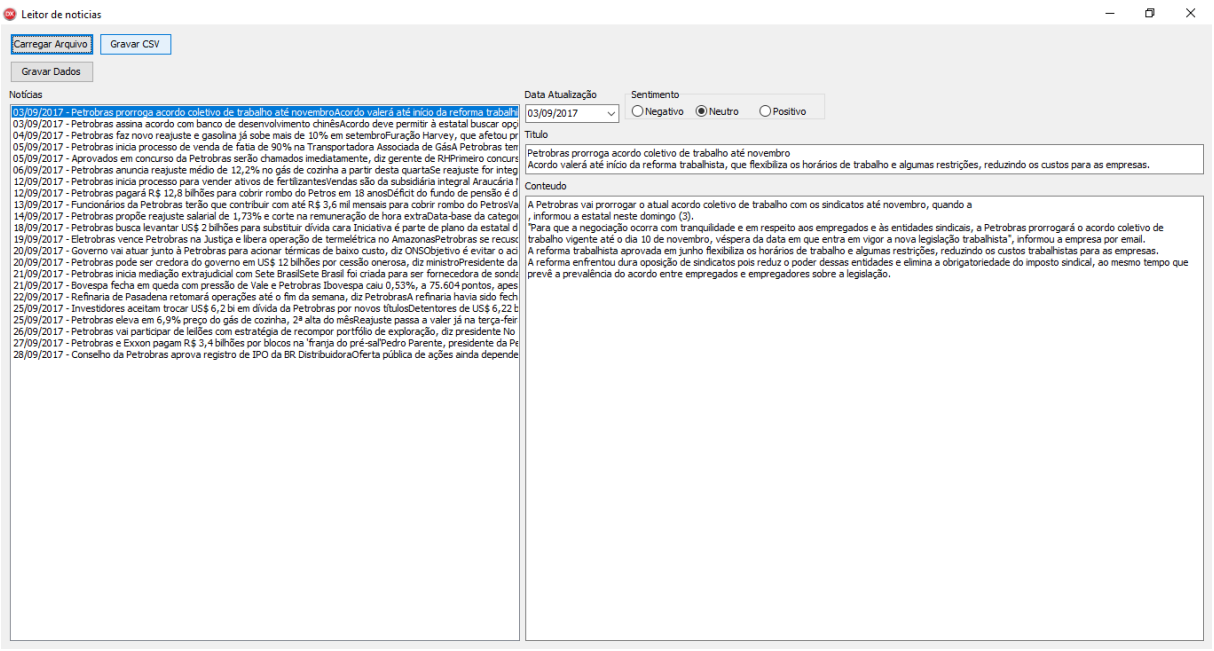

Figura 6: Programa de rotulação dos dados

Para o desenvolvimento do rotulador, foi utilizado a linguagem Delphi, pela velocidade de desenvolvimento das telas, nele lemos o arquivo JSON, ou o CSV, que foi baixado da plataforma Scrapy Cloud, então decidimos mostrar todas as notícias ordenadas pela data no lado esquerdo, e ao selecionar um item os dados dela são mostrados no lado direito da tela, entre os dados está o dado de sentimento, no qual pode ser negativo, neutro ou positivo, podemos alterar qualquer um dos dados do item, e após isso salvar as mudanças, após a alteração de todos os dados necessários salvamos as mudanças em um arquivo CSV.

3.2.4 Análise de sentimento das notícias

3.2.4.1 Pré-Processamento

Após a etapa de extração de notícias, obtemos arquivos *CSV*, e precisamos limpar os dados para nosso uso, começamos deixando somente as letras, deixando todo nosso texto em minúsculo, após isso em nossa análise *lexicon* também fizemos a tradução de nossos dados para o inglês, e removemos os *stopwords*.

3.2.4.2 Lexicon

Inicialmente, para a análise de sentimento utilizamos a abordagem *lexicon*, que é uma abordagem mais simples, para isso utilizamos a biblioteca *TextBlob*, tendo como objetivo buscar a polaridade de nosso documento.

O primeiro método aplicado foi a divisão do documento em uma lista de palavras, após isso vimos a polaridade de cada uma das palavras, porém não obtivemos um bom resultado, após isso tentamos utilizar o documento divido por sentenças, porém continuamos com um resultado insatisfatório, após isso tentamos verificar a polaridade do documento como um todo, porém os resultados continuaram o mesmo.

Analisando os resultados, decidimos utilizar o método de aprendizado de máquina, pois as notícias são muito subjetivas, logo analisar a polaridade das palavras não irá nos levar num resultado satisfatório.

3.2.4.3 Aprendizado de máquina

Utilizando a biblioteca *scikit-learn*, em conjunto com outras bibliotecas, tais como a *TextBlob*, e a *NLTK*. construímos um modelo, para classificar os sentimentos entre positivo, neutro ou negativo. Decidimos utilizar o modelo de Bayes, o modelo *sklearn.naive_bayes.MultinomialNB* para ser mais exato.

Primeiramente foi necessário treinar nosso modelo, para isso fizemos o pré-processamento dos dados rotulados anteriormente, referentes ao período de maio a agosto, criamos um *bag of words* e criamos um modelo de espaço vetorial com este *bag of words*, para conseguirmos obter nossas *features*, após isso treinamos nosso classificador com este modelo vetorial. Utilizamos o mesmo processo para nossos dados do período de setembro, e com isso passamos o modelo

vetorial do período de setembro para conseguirmos previsões de nossos dados, com isso salvamos estas previsões em um arquivo CSV, para a utilização posterior.

3.2.5 Correlação dos dados

A correlação mede o grau no qual duas variáveis tendem a mudar juntas. Este coeficiente mostra a força e a direção deste relação, neste projeto foi decido o uso do método de correlação de pearson.

O coeficiente de correlação de Pearson mede a relação linear entre dois conjuntos de dados. Estritamente falando, a correlação de Pearson exige que cada conjunto de dados seja normalmente distribuído. Como outros coeficientes de correlação, este varia entre -1 e +1 com 0, o que implica que não há correlação. As correlações de -1 ou +1 implicam uma relação linear exata. As correlações positivas implicam que, à medida que x aumenta, o mesmo acontece com y. As correlações negativas implicam que, como x, aumenta, e diminui. (SCIPY, 2017)

O valor p indica aproximadamente a probabilidade de um sistema não correlacionado produzir conjuntos de dados que tenham uma correlação de Pearson pelo menos tão extrema como a calculada a partir desses conjuntos de dados. Os valores de p não são inteiramente confiáveis, mas provavelmente são razoáveis para conjuntos de dados maiores que 500 ou mais. (SCIPY, 2017)

Para a correlação dos dados decidimos utilizar a biblioteca *scipy*, utilizamos a função *scipy.stats.pearsonr*, que calcula o coeficiente de correlação de Pearson, e o valor p, para isso precisamos passar um array x e, um array y, em nosso caso eles seriam respectivamente, os sentimentos classificados, e os valores das ações.

Decidimos fazer a correlação entre o sentimento de uma dia X e verificar a correlação deste sentimento no mesmo dia, assim como nos próximos dias transcorridos, fizemos o teste utilizando até 7 dias transcorridos, agrupando os dados por meses, e também fazendo um agrupamento de todos os meses.

4 Resultados

4.1 Importação dos dados de ações

Como visto anteriormente foi feito o download de um arquivo com o histórico das ações da BM&FBOVESPA, e transferidos somente os dados referentes a Petrobras para o banco de dados, os 10 primeiros registro do CSV pode ser vistos na figura 7, e os 10 primeiros registros no banco de dados podem ser visto na tabela 1 a titulo de ilustração

```
00COTAHIST.2017BOVESPA 20171004
012017010202AALR3       010ALLIAR      ON      NM   R$  000000000146200000000014880000000001440000000000145800000000014600000000
012017010210ABCB2       010ABC BRASIL  DIR PRE N2   R$  000000000028500000000002850000000002850000000002850000000002850000000
012017010202ABCB4       010ABC BRASIL  PN  EJS N2   R$  000000000134000000000013520000000013080000000001330000000001331000000
012017010296ABCB4F      020ABC BRASIL  PN  EJS N2   R$  000000000134500000000013480000000013150000000001329000000000132600000000
012017010212ABCP11      010FII ABC IMOBCI  ER       R$  000000000011830000000001183000000000110100000000011620000000001136000000
012017010202ABEV3       010AMBEV S/A   ON  EJ       R$  000000000163400000000016660000000016260000000001642000000000163100000000
012017010296ABEV3F      020AMBEV S/A   ON  EJ       R$  000000000164100000000016660000000016270000000001645000000000166000000000
012017010262ABEV3T      030AMBEV S/A   ON  EJ    016R$  000000000165500000000016560000000016490000000001653000000000165000000000
012017010262ABEV3T      030AMBEV S/A   ON  EJ    045R$  000000000165600000000016570000000016560000000001656000000000165700000000
012017010262ABEV3T      030AMBEV S/A   ON  EJ    060R$  000000000169100000000016940000000016910000000001692000000000169400000000
```

Figura 7: Arquivo bovespa

Data pregão	Código BDI	Código Negociação	Nome Resumido	Moeda	Preço Ultimo
20170102	02	PETR4	PETROBRAS	R$	14.66
20170102	96	PETR4F	PETROBRAS	R$	14.65
20170102	62	PETR4T	PETROBRAS	R$	14.75
20170102	62	PETR4T	PETROBRAS	R$	14.82
20170102	62	PETR4T	PETROBRAS	R$	14.95
20170102	62	PETR4T	PETROBRAS	R$	15.11
20170103	02	PETR4	PETROBRAS	R$	15.5
20170103	96	PETR4F	PETROBRAS	R$	15.44
20170103	62	PETR4T	PETROBRAS	R$	15.69
20170103	62	PETR4T	PETROBRAS	R$	15.58

Tabela 1: Registro SQL

Os dados do arquivo figura 7 seguem um layout que pode ser encontrado no site da BM&FBOVESPA, dentro desse layout a primeira linha do arquivo é o "REGISTRO - 00 - HEADER", e as outras linhas o "REGISTRO - 01 - COTAÇÕES HISTÓRICAS POR PAPEL-MERCADO", no caso importamos todos os campos do Registro01 para o banco de dados, porém em nosso projeto usaremos apenas 3 campos: a data do pregão, o código de negociação e o preço ultimo, a data do pregão e o código negociação foram utilizadas para o filtro onde a data do pregão é a data de referência da ação e o código negociação sempre será filtrado por PETR4 que são as ações que estamos utilizando em nosso projeto, e o preço ultimo é o campo que será usado como valor referência da ação no dia.

4.2 Extração de notícias

No processo de extração de noticias foram gerados dois tipos de arquivo, o tipo CSV, e JSON, os 10 primeiros registros do arquivo CSV podem ser vistos na tabela 2 e um registro do arquivo JSON pode ser visto na figura 8

"_type"	"conteudo"	"data_atualizacao"	"titulo"
"NoticiasItem"	", O ministro de Minas e Energia, ..."	"20/09/2017 18h15 "	"Petrobras pode ser credora ..."
"NoticiasItem"	", Detentores de US$ 6,22 bilhões em ..."	"25/09/2017 11h03 "	"Investidores aceitam trocar ..."
"NoticiasItem"	", Os lances bilionários oferecidos ..."	"27/09/2017 15h26 "	"Petrobras e Exxon pagam R$ ..."
"NoticiasItem"	", A Petrobras deverá gastar R$ 12,8 ..."	"12/09/2017 20h31 "	"Petrobras pagará R$ 12,8 ..."
"NoticiasItem"	", A subsidiária de geração da estatal ..."	"19/09/2017 13h48 "	"Eletrobras vence Petrobras ..."
"NoticiasItem"	", A Petrobras informou nesta terça ..."	"05/09/2017 20h43 "	"Petrobras inicia processo de ..."
"NoticiasItem"	", A Petrobras iniciou processo para a ..."	"12/09/2017 06h59 "	"Petrobras inicia processo para ..."
"NoticiasItem"	", A Petrobras informou nesta quinta ..."	"21/09/2017 19h12 "	"Petrobras inicia mediação ..."
"NoticiasItem"	", Os concursos ,, com possibilidade de , ..."	"05/09/2017 06h00 "	"Aprovados em concurso ..."
"NoticiasItem"	", A Petrobras planeja levantar US$ 2 ..."	"18/09/2017 13h06 "	"Petrobras busca levantar ..."

Tabela 2: Arquivo CSV extraido

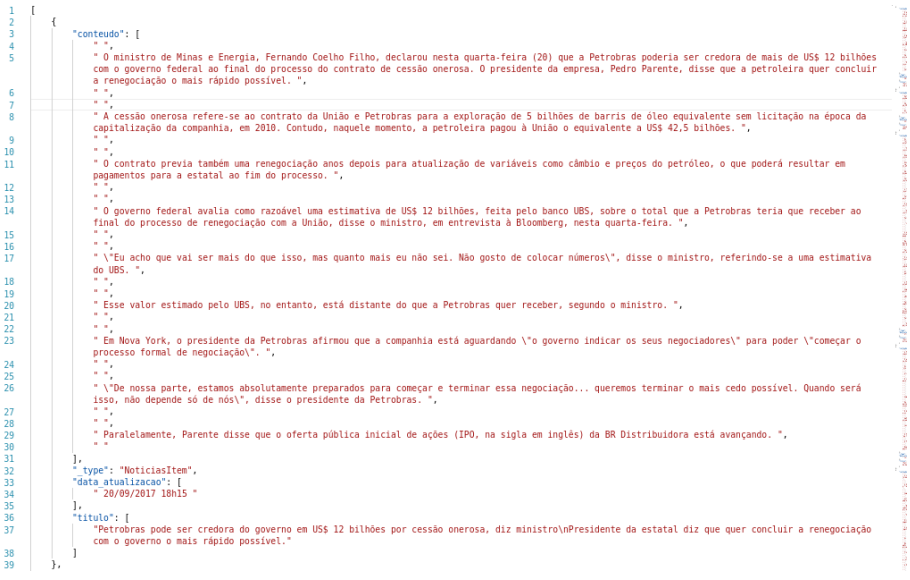

Figura 8: Arquivo JSON extraido

Observando os dados do arquivo CSV na tabela 2, verificamos que ele precisa de algumas pequenas correções como as aspas duplas nos campos, e o campo data_atualizacao segue um formato de data incorreto do que desejamos que seria seguindo o formato "dd/mm/yyyy", já o arquivo JSON da figura 8 segue um padrão incomum do JSON, onde os campos de string estão vindo como uma lista de string.

4.3 Rotulação dos dados

O processo de rotulação de dados gera um arquivo do tipo CSV e seus 10 primeiros registros podem ser ser observados na tabela 3, observando esse arquivo verificamos que as correções necessárias do arquivo extraído CSV mostrado na tabela 2 foi realizado.

data_atualizacao	titulo	conteudo	sentimento
02/05/2017	Justiça libera venda de área ...	A Petrobras informou nesta terça-feira (2) que foi suspens ...	1
02/05/2017	Presidente da Petrobras diz q...	O fraco desempenho dos preços do petróleo neste ano não de ...	1
03/05/2017	Decreto regulamenta direito d...	Um decreto publicado no Diário Oficial da União nesta quar ...	-1
08/05/2017	Representante da Petrobras na...	O representante local da Petrobras da Bolívia está em pris ...	-1
08/05/2017	Justiça extingue ação contra ...	A Justiça Federal de Sergipe atendeu a pedido da Petrobras ...	1
09/05/2017	Petrobras diz que ainda não e...	A Petrobras informou nesta terça-feira (9) em comunicado q ...	0
10/05/2017	Bovespa tem forte alta puxada...	O principal índice da bolsa paulista opera em alta nesta q ...	1
10/05/2017	Petrobras coloca refinaria de...	A Petrobras incluiu na sua carteira de ativos à venda a re ...	-1
11/05/2017	Petrobras tem lucro de R$ 445...	A Petrobras registrou lucro de R$ 445 bilhões no 1° trim i...	1
11/05/2017	Bovespa fecha em alta à esper...	O principal índice da bolsa paulista fechou em alta nesta ...	1

Tabela 3: Dados rotulados

4.4 Análise de sentimento das noticias

Os registros dos dados da análise de sentimento obtidos em nosso projeto pela abordagem de léxico podem ser observados na tabela 4.

data	titulo	sentimento	polaridade
02/10/2017	Petrobras anuncia US$ 63 bi em operações de pagamento renegociação e contratação ...	18.80%	46.61%
02/10/2017	Ministro de Minas e Energia prevê que Petrobras será privatizada: É um caminho F ...	-4.17%	46.46%
03/10/2017	Não estamos tratando disso diz ministro sobre privatização da Petrobras Fernando ...	-1.81%	31.11%
04/10/2017	Comitê do setor elétrico pedirá que Petrobras forneça combustível para térmicas ...	6.50%	40.25%
04/10/2017	Petrobras está perto de atingir meta de desalavancagem A companhia que tem reduz ...	0.31%	29.06%
05/10/2017	Indústria naval deve recorrer de decisão da ANP que libera Petrobras para constr ...	5.37%	36.70%
05/10/2017	Petrobras vai analisar pedido do governo por ajuda a térmicas mas sem subsídio O ...	6.31%	32.82%
05/10/2017	Petrobras vê como positiva flexibilização de conteúdo local em reserva do pré-sa ...	3.21%	19.23%
10/10/2017	Petrobras eleva preço do botijão de gás em 129% a partir desta quarta Estatal es ...	-1.20%	45.94%
11/10/2017	Cade vê como complexa compra de ativos da Petrobras por mexicana e pede estudos ...	0.00%	22.50%
11/10/2017	TCU bloqueia bens de Dilma por prejuízo à Petrobras com compra de Pasadena Além ...	3.36%	31.16%
16/10/2017	Petrobras pede registro de companhia aberta para BR Distribuidora Pedido de IPO ...	-8.57%	34.29%
17/10/2017	Moodys eleva nota da Petrobras e muda perspectiva para estável Agência cita redu ...	11.77%	41.30%
18/10/2017	Produção da Petrobras no Brasil sobe 28% em setembro ante agosto Na comparação c ...	-2.45%	37.00%
19/10/2017	Governo exclui fatia da Petrobras na Braskem de programa de desestatização Medid ...	0.00%	24.00%
19/10/2017	Carf decide a favor da Petrobras em processo de R$ 78 bilhões diz estatal Estata ...	1.11%	2.22%
24/10/2017	Petrobras quer vender a BR Distribuidora 'o mais rápido possível' Empresa tem me ...	0.81%	46.20%
24/10/2017	Petrobras fará parceria para disputar blocos do pré-sal em leilão na sexta-feira ...	12.25%	55.80%
26/10/2017	Petrobras aprova adesão a Nível 2 de governança da bolsa e corte de gerências Pa ...	12.53%	45.32%
26/10/2017	Petrobras pode incluir ativos de logística em vendas de refinarias; detalhes dev ...	17.10%	54.89%
26/10/2017	Petrobras será 'seletiva' mas 'muito firme' no leilão do pré-sal diz Pedro Paren ...	18.77%	52.49%
26/10/2017	Petrobras adere ao programa de regularização de dívida não tributária Segundo a ...	2.90%	39.61%
27/10/2017	Petrobras 'não podia se dar ao luxo de perder essa oportunidade' diz Parente sob ...	12.41%	38.96%
27/10/2017	Com lances altos Petrobras leva 3 blocos do pré-sal e oferece até 80% da produçã ...	11.45%	48.39%
31/10/2017	Petrobras busca parcerias para concluir obras do Comperj diz Parente Presidente ...	7.20%	37.36%
31/10/2017	Parceria da Petrobras com empresa britânica BP deve envolver troca de ativos diz ...	10.94%	43.05%

Tabela 4: Sentimento lexicon

Com os dados mostrados na análise *lexicon* conseguimos observar que em sua maioria os sentimentos das notícias tendem a zero e sua subjetividade é alta, estes dados em nosso projeto não seriam relevantes pois as ações não seguem uma variação muito sutil, precisamos de sentimentos de notícias com variações maiores para analisarmos sua correlação.

Os registros dos dados da análise de sentimento para o mês de outubro com a abordagem do aprendizado de máquina podem ser observados na tabela 5, além de métricas relevantes ao modelo.

data	titulo	sentimento
02/10/2017	Petrobras anuncia US$ 63 bi em operações de pagamento renegociação e contratação de dívida Empresa pagou antec ...	1
02/10/2017	Ministro de Minas e Energia prevê que Petrobras será privatizada: É um caminho Fernando Coelho Filho afirmou n ...	1
03/10/2017	Não estamos tratando disso diz ministro sobre privatização da Petrobras Fernando Coelho Filho (Minas e Energia ...	1
04/10/2017	Comitê do setor elétrico pedirá que Petrobras forneça combustível para térmicas paradas Comitê destacou que a ...	1
04/10/2017	Petrobras está perto de atingir meta de desalavancagem A companhia que tem reduzido trimestre a trimestre sua ...	1
05/10/2017	Indústria naval deve recorrer de decisão da ANP que libera Petrobras para construir casco de Libra no exterior ...	0
05/10/2017	Petrobras vai analisar pedido do governo por ajuda a térmicas mas sem subsídio O Comitê de Monitoramento do Se ...	1
05/10/2017	Petrobras vê como positiva flexibilização de conteúdo local em reserva do pré-sal Na véspera ANP flexibilizou ...	0
10/10/2017	Petrobras eleva preço do botijão de gás em 129% a partir desta quarta Estatal estima que preço ao consumidor f ...	0
11/10/2017	Cade vê como complexa compra de ativos da Petrobras por mexicana e pede estudos Venda da Petroquímica Suape e ...	1
11/10/2017	TCU bloqueia bens de Dilma por prejuízo à Petrobras com compra de Pasadena Além da ex-presidente decisão ating ...	-1
16/10/2017	Petrobras pede registro de companhia aberta para BR Distribuidora Pedido de IPO inclui também aval para realiz ...	1
17/10/2017	Moodys eleva nota da Petrobras e muda perspectiva para estável Agência cita redução do endividamento da petrol ...	1
18/10/2017	Produção da Petrobras no Brasil sobe 28% em setembro ante agosto Na comparação com setembro de 2016 houve qued ...	-1
19/10/2017	Governo exclui fatia da Petrobras na Braskem de programa de desestatização Medida também exclui a fatia da pet ...	-1
19/10/2017	Carf decide a favor da Petrobras em processo de R$ 78 bilhões diz estatal Estatal divulgou comunicado nesta qu ...	1
24/10/2017	Petrobras quer vender a BR Distribuidora 'o mais rápido possível' Empresa tem meta de vender ativos que somam ...	1
24/10/2017	Petrobras fará parceria para disputar blocos do pré-sal em leilão na sexta-feira Presidente da estatal adianto ...	0
26/10/2017	Petrobras aprova adesão a Nível 2 de governança da bolsa e corte de gerências Patamar intermediário de governa ...	1
26/10/2017	Petrobras pode incluir ativos de logística em vendas de refinarias; detalhes devem sair no 1° tri diz Parente ...	0
26/10/2017	Petrobras será 'seletiva' mas 'muito firme' no leilão do pré-sal diz Pedro Parente ANP pôs à venda 8 blocos de ...	0
26/10/2017	Petrobras adere ao programa de regularização de dívida não tributária Segundo a estatal adesão resultará em im ...	1
27/10/2017	Petrobras 'não podia se dar ao luxo de perder essa oportunidade' diz Parente sobre alta oferta nos leilões Est ...	0
27/10/2017	Com lances altos Petrobras leva 3 blocos do pré-sal e oferece até 80% da produção à União No regime de partilh ...	0
31/10/2017	Petrobras busca parcerias para concluir obras do Comperj diz Parente Presidente da estatal revelou a expectati ...	1
31/10/2017	Parceria da Petrobras com empresa britânica BP deve envolver troca de ativos diz Parente Estatal divulgou cart ...	1

Tabela 5: Sentimento scikit

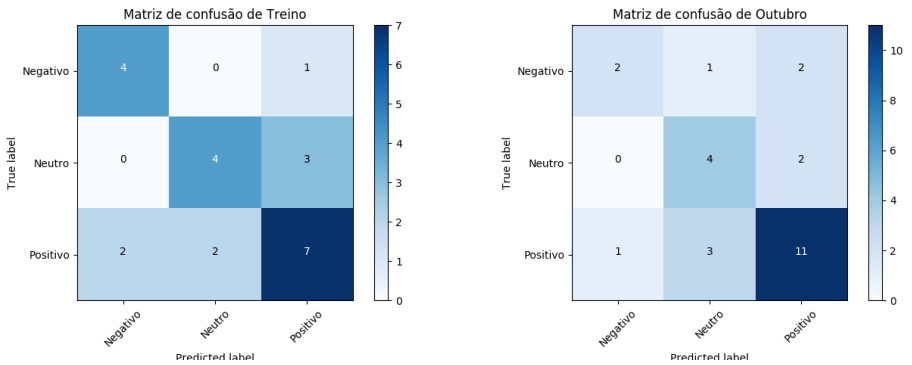

Figura 9: Matriz de confusão das métricas

Fonte dos dados	Kappa	Precisão	Recall	FScore
Treino	45.24%	65.66%	66.93%	65.97%
Outubro	39.69%	63.33%	60.00%	60.16%

Tabela 6: Métricas do modelo

Observando as métricas obtidas através da matriz de confusão percebemos que a precisão dos dados obtidas pelos dados de treino e pelo mês de outubro são razoáveis, em nosso caso de treino obtivemos uma precisão de 65% com um Kappa de 45%, e no mês de outubro obtivemos uma precisão de 63% com um Kappa de 39%, nossa métrica Kappa nos indica uma concordância entre moderada e razoável, e nossa precisão também pode ser vista como razoável em ambos os casos, tentamos melhorar esses dados porem não conseguimos, achamos que isso ocorre devido ao alto grau de subjetividade dos dados.

4.5 Correlação dos sentimentos com as ações

Os dados da correlação de Pearson e os gráficos relevantes ao mesmo podem ser observados abaixo na tabela 7, figura 11 e figura 10 respectivamente.

Fonte Dados	Dias transcorrido	Correlação	pvalue
Dados de maio	0	16.96%	53.00%
Dados de junho	0	3.91%	89.45%
Dados de julho	0	-9.19%	76.53%
Dados de agosto	0	4.62%	89.92%
Dados de setembro	0	-16.37%	54.46%
Dados de outubro	0	22.58%	43.77%
Dados da previsão	0	-14.97%	60.95%
Dados de maio	1	45.98%	7.32%
Dados de junho	1	4.50%	87.86%
Dados de julho	1	23.81%	43.34%
Dados de agosto	1	19.04%	59.84%
Dados de setembro	1	-33.18%	20.92%
Dados de outubro	1	43.21%	12.29%
Dados da previsão	1	15.71%	59.16%
Dados de maio	2	43.17%	9.50%
Dados de junho	2	22.21%	44.53%
Dados de julho	2	14.32%	64.08%
Dados de agosto	2	51.83%	12.48%
Dados de setembro	2	-33.38%	20.64%
Dados de outubro	2	32.28%	26.03%
Dados da previsão	2	-0.86%	97.67%
Dados de maio	3	41.98%	10.54%
Dados de junho	3	25.28%	38.33%
Dados de julho	3	10.16%	74.11%
Dados de agosto	3	28.07%	43.21%
Dados de setembro	3	-28.33%	28.77%
Dados de outubro	3	42.60%	12.89%
Dados da previsão	3	9.12%	75.65%
Dados de maio	4	37.04%	15.78%
Dados de junho	4	30.48%	28.93%
Dados de julho	4	6.67%	82.86%

Fonte Dados	Dias transcorrido	Correlação	pvalue
Dados de agosto	4	33.48%	34.43%
Dados de setembro	4	-25.16%	34.72%
Dados de outubro	4	37.30%	18.90%
Dados da previsão	4	13.25%	65.16%
Dados de maio	5	38.57%	14.01%
Dados de junho	5	26.69%	35.63%
Dados de julho	5	4.07%	89.51%
Dados de agosto	5	1.26%	97.25%
Dados de setembro	5	-7.52%	78.21%
Dados de outubro	5	43.15%	12.34%
Dados da previsão	5	11.71%	69.01%
Dados geral	0	18.12%	10.12%
Dados geral	1	24.52%	2.55%
Dados geral	2	24.55%	2.53%
Dados geral	3	24.68%	2.45%
Dados geral	4	24.75%	2.41%
Dados geral	5	25.21%	2.15%
Dados geral	6	22.01%	4.56%
Dados geral	7	17.95%	10.44%

Tabela 7: Correlação de sentimento vs valor da ação - Normalizado

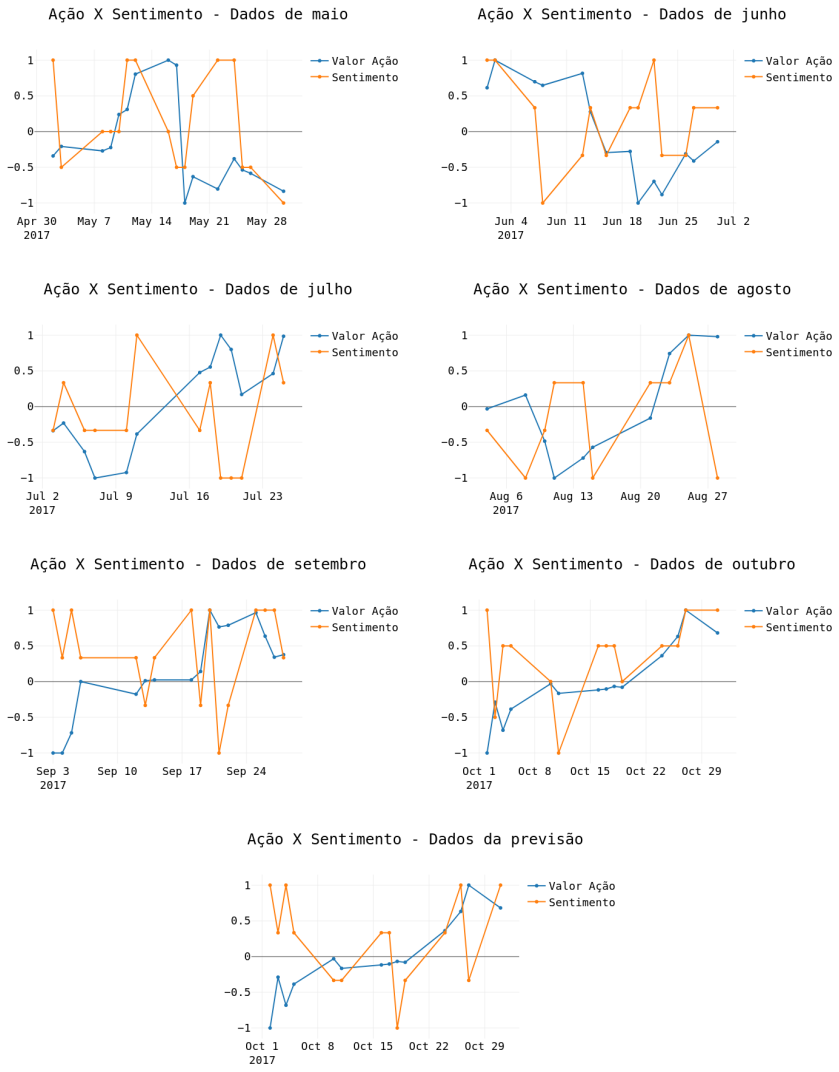

Figura 10: Gráficos de Ação X Sentimento - Normalizado

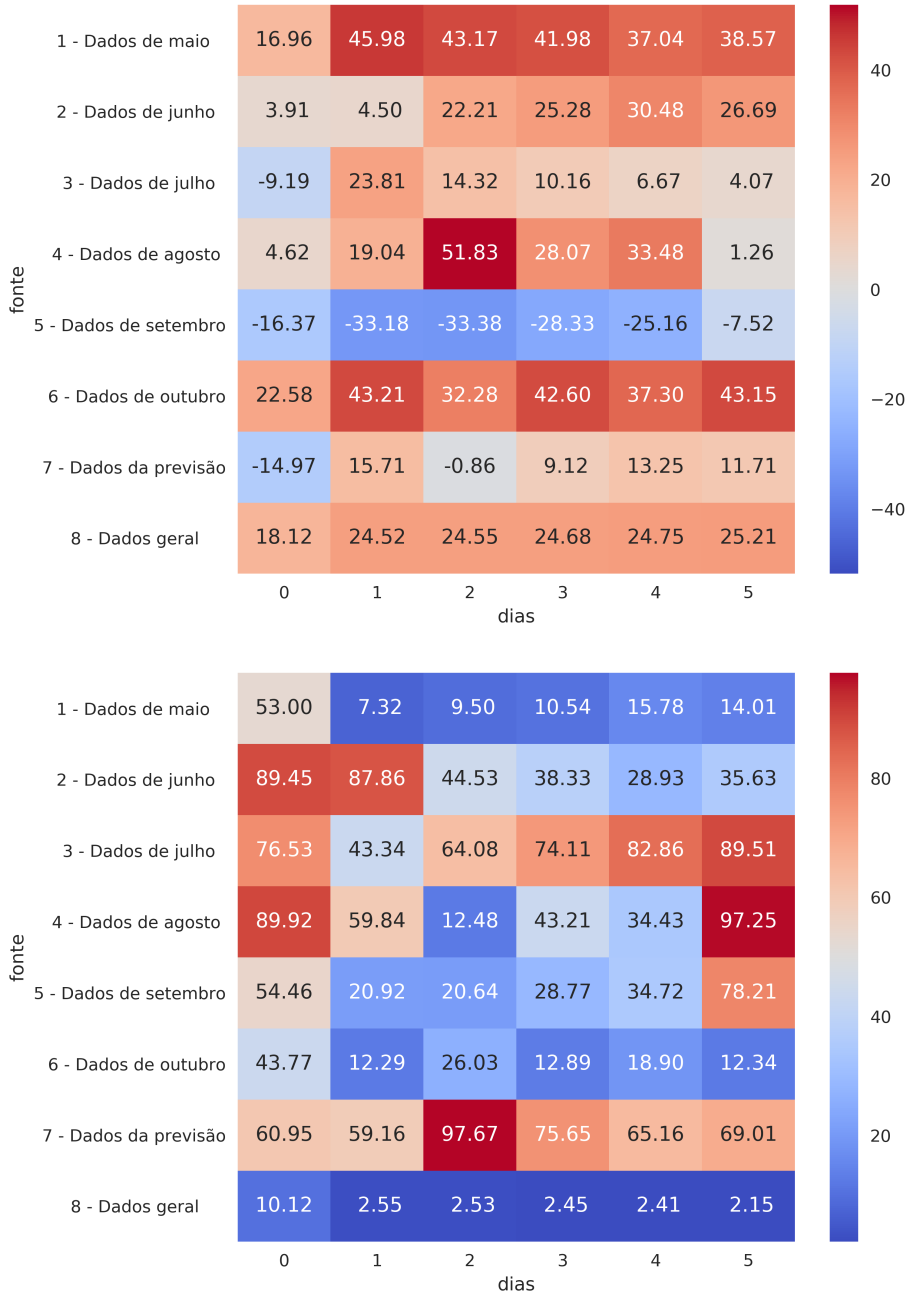

Figura 11: Gráficos de mapas de calor dos campos de correlação e p-value

Os dados mostrados acima foram feitos utilizando os dados rotulados manualmente, e também utilizando os dados da previsão para o mês de outubro. Como citado anteriormente os testes foram feitos agrupando os dados por meses de maio a outubro, e também agrupando todos os dados de maio a outubro, verificando os sentimentos com os respectivos dias transcorridos como referência.

Analisando os dados apresentados vimos que os sentimentos de maneira geral possuem uma correlação baixa com os valores da bolsa de ações. Porém alguns meses como maio e outubro possuem uma correlação superior. Os limites das correlações vão de 51.83% a -33.38%.

Conseguimos observar também que se agruparmos todos os dados temos um correlação baixa, porém relevante com o valor médio de 20%, começando com um valor de 18% com 0 dias transcorridos e chegando em seu pico a 25% de correlação após 5 dias transcorridos, verificando os valores de *p-value*, e assumindo um nível de significância de 5% como referência conseguimos observar que com os dados de todos meses agrupados conseguimos um valor bom de 2.15% que é menor do que nosso valor de referência 5%. Verificando os gráficos conseguimos perceber que esses valores fazem algum sentido pois conseguimos perceber alguns padrões onde aparenta-se que os sentimentos impactam os valores das ações após alguns dias. Conseguimos também observar como os dados de nossa previsão e compara-los com os dados do mês de referência que seria o de outubro, verificando assim uma divergencia bem alta entre os dois, algo esperado levando a precisão de aproximadamente 60% da classificação do sentimentos.

Levando em conta todos os fatores analisados supomos que para melhorar o valor da correlação encontrada deveríamos considerar mais parâmetros, e não somente os sentimentos das noticias do dia, além disso precisaríamos de dados de mais fontes e não somente do site da G1. Após isso também fizemos algumas pesquisas e foi suposto que algumas notícias podem ter um impacto maior que outras e isso não foi levado em consideração ao fazer nossas correlações, um exemplo que pode ser citado foi uma notícia de setembro do site da Exame que informa as 10 melhores ações para se investir em setembro segundo 18 corretoras, e dentre estas a ação da Petrobras estava em primeiro e, após essa notícia os valores da Petrobras tiveram um resultado positivo de maneira geral, neste sentido, verifica-se um maior impacto de uma notícia provinda de outra fonte, que não a utilizada no presente trabalho.

5 Desafios e Dificuldades

No decorrer de nosso projeto foram encontradas várias dificuldades a primeira que podemos citar foi a dificuldade em encontrar dados referentes a sentimentos de notícias rotulados, procuramos em vários sites, porém o único lugar que foi encontrado algo do gênero foi no site da kaggle um *dataset* que possuía as 25 notícias mais populares do dia no site reddit, a data e o sentimento. Porém não conseguimos aproveitar os dados.

A falta de um dataset nos levou a extração de notícias onde também encontramos dificuldades, a principal dele foi que precisamos filtrar os dados entre um período e de uma fonte de forma prática, conseguimos resolver isso com a pesquisa do google, porém regras configuradas no arquivo *robot.txt* do site do google, bloqueavam nossa extração após uma certa quantidade de dados que extraiamos conseguimos resolver isso após utilizarmos a plataforma *ScrapingHub*.

Após conseguirmos a extração dos dados encontramos outro problema que foi a rotulação dos dados, não conseguimos encontrar um modo prático para a resolução deste problema, então criamos um software para ajudar na rotulação dos dados, e rotulamos os dados manualmente.

No decorrer do projeto foi levantado que os dados não são adequados para a análise de sentimento, pois contem muita subjetividade, o que muitas vezes dificultam a classificação de um sentimento até mesmo para as pessoas que estão rotulando os dados.

6 Conclusão e Trabalhos Futuros

6.1 Conclusão

O mercado de ações é algo muito complexo e volátil, e que leva muitas fatores em conta, um desses fatores pode ser encontrado na análise fundamentalista, onde os foco são fatores externos como notícias para a previsão dessa volatilidade, e é nesse ponto que o objetivo geral deste trabalho foca, criando um classificador de sentimento de notícias a fim de se obter a correlação delas com os valores das ações.

Em ao objetivo específico "Extração de notícias de sites", foi feito uma pesquisa com as técnicas para a extração de dados de sites da internet, assim como suas ferramentas, foram escolhido a linguagem Python em conjunto com o *framework* Scrapy, porém enfrentamos alguns problemas e tivemos que usar também a plataforma ScrapyCloud da empresa ScrapingHub.

Para a "Análise de sentimento das notícias extraídas", foi feito uma estudo sobre as técnicas de classificação de sentimento, e primeiramente decidimos utilizar a técnica do *lexicon* em conjunto com uma ferramenta Python chamada TextBlob, porém após resultados não satisfatórios e mais estudos, decidimos usar o aprendizado de maquina para a solução do problema, utilizando a ferramenta scikit-learn chegamos a resultados razoáveis com uma precisão de 63.33% e um Kappa de 39.69%.

A "Análise da correlação dos dados da análise de sentimento com os dados de valorização ou desvalorização das mesmas", foi feita utilizando a biblioteca SciPy, para está análise foi preciso importar os dados da ações da BM&FBOVESPA, utilizamos a biblioteca Peewee em para armazenar os dados em um banco de dados SQLite, com isso fizemos a análise dos valores das ações da Petrobras com os sentimentos rotulados para o treino da análise de sentimento, e com os valores da previsão do mês de outubro, os resultados mostraram uma correlação consistente, porém baixa com os dados de todos os meses agrupados com uma média 20%.

Após atingidos os objetivos específicos deste trabalho com base na análise dos resultados obtidos, pôde-se concluir que os sentimentos de notícias são algo muito subjetivos, porém os sentimentos destas notícias podem impactar os valores das ações mesmo que sutilmente em nosso contexto.

6.2 Trabalhos Futuros

1. Utilização de filtros mais abrangentes, como a utilização de mais empresas como referência, assim como um período maior.
2. Extração de dados de mais fontes como a Exame, InfoMoney, entre outras fontes.
3. Rotulação dos dados por mais de 1 pessoa.
4. Verificação da correlação de dados dos dados com as ações utilizando outras fontes de dados além do sentimento, como dados financeiros da empresa entre outros.
5. Utilização de outros modelos, técnicas e ferramentas para a classificação do sentimento.
6. Utilização de uma base de dados diferentes para a correlação, um exemplo seria utilizar os comentários das noticias.
7. Utilização de pesos para os sentimentos das noticias mais importantes.

Referências

ALVES, D. S. Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores. *Universidade de Brasília,(2015 http://repositorio.unb.br/handle/10482/19345)*, 2015.

AULERIO. *Dicionario Aurelio*. 2017. Disponível em: <https://dicionariodoaurelio.com/>.

BARBER, D. *Bayesian Reasoning and Machine Learning*. [S.l.]: Cambridge University Press, 2016.

BERTOLO, L. *Mercado Financeiro*. [s.n.], 2002. Disponível em: <http://lbertolo.tripod.com/MATEMATICAFINANCEIRA.pdf>.

FISCHER, A. et al. *Mercado de Valores Mobiliários Brasileiro*. Macgraw-Hill, 2014. Disponível em: <www.cvm.gov.br/menu/investidor/publicacoes/livros.html>.

FONSECA, J. W. F. da. *Análise e Decisão de Investimentos*. [S.l.]: IESDE Brasil, 2009.

HU, M.; LIU, B. *Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014. Disponível em: <http://doi.acm.org/10.1145/1014052.1014073>.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. MIT Press, 2009. Disponível em: <http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105975>.

LIU, B. *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers, 2012. Disponível em: <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>.

MITTAL, A.; GOEL, A. Stock prediction using twitter sentiment analysis. *Stanford University, CS229 (2011 http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf)*, v. 15, 2012.

MONARD, M. C.; BARANAUSKAS, J. A. Sistemas inteligentes: fundamentos e aplicações. In: _____. Manole ltda, 2003. cap. 4. Disponível em: <http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>.

MURPHY, J. J. *Technical Analysis Of The Financial Markets*. [S.l.: s.n.], 1999. 1 – 239 p. ISSN 00314439. ISBN 0735200661 9780735200661.

NUIJ, W. et al. An automated framework for incorporating news into stock trading strategies. *IEEE Transactions on Knowledge and Data Engineering*, v. 26, n. 4, p. 823–835, 2014. ISSN 10414347.

POOLE, D. L.; MACKWORTH, A. K. *Artificial Intelligence: foundations of computational agents*. [S.l.]: Cambridge University Press, 2010.

RUSSELL, S.; NORVIG, P. *Inteligência Artificia*. 3. ed. [S.l.]: Elsevier, 2013.

SCIPY. *Scipy | scipy.stats.pearsonr*. 2017. Disponível em: <https://docs.scipy.org/doc/scipy-0.19.1/reference/generated/scipy.stats.pearsonr.html>.

SCRAPY. *Scrapy | A Fast and Powerful Scraping and Web Crawling Framework*. 2017. Disponível em: <https://scrapy.org/>.

SELAN, B. *Mercado Financeiro*. Universidade Estácio de Sá, 2015. Disponível em: <https://profhubert.yolasite.com/resources/LIVRO\%20PROPRIETARIO\%20-\%20Mercado\%20financeiro.pdf>.

TSYTSARAU, M.; PALPANAS, T. *Survey on mining subjective data on the web.Data Mining and Knowledge Discovery*. MIT Press, 2012. Disponível em: <http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105975>.

Universidade Estadual de Maringá
Departamento de Informática
Av. Colombo 5790, Maringá-PR, CEP 87020-900
Tel: (44) 3261-4324 Fax: (44) 3263-5874
www.din.uem.br