https://eventos.utfpr.edu.br//sicite/sicite2020

Análise de sentimento em tempo real de notícias do mercado de ações

Real-time sentiment analysis of stock market news

Vinicius Augusto de Souza
vsouza.1998@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Giovani Volnei Meinerz
giovanimeinerz@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

RESUMO

Este artigo descreve o treinamento de um classificador Naive Bayes, baseado em tweets previamente classificados quanto a sua polaridade, o qual é utilizado para realizar uma análise de sentimento sobre textos gerados em *stream*, provenientes da rede social Twitter. O objetivo deste estudo é realizar uma análise de sentimento em tempo real, sobre *tweets* publicados por veículos de notícias especializados no mercado de ações brasileiro. O desenvolvimento do analisador de sentimento requereu 6 fases: armazenamento da base de *tweets* manualmente rotulados; pré-processamento, utilizando métodos de limpeza do texto, tokenização e remoção de *stopwords*; aplicação do método *Term Frequency–Inverse Document Frequency* (TF-IDF); treinamento do classificador; coleta dos novos *tweets* em *streaming* e; análise de sentimento em tempo real. Após o desenvolvimento das fases citadas anteriormente, o analisador de sentimento atingiu uma acurácia de 76,8 por cento.

PALAVRAS-CHAVE: Aprendizado do Computador. Processamento de linguagem natural. Twitter.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

ABSTRACT

This article describes the training of a Naive Bayes classifier, based on a previously classified base of tweets as to their polarity, which is used to perform a sentiment analysis on texts generated in streams, coming from the social network Twitter. The objective of this study is to carry out an analysis of sentiment in real time, on tweets published by news outlets specialized in the brazilian stock market. The development of the sentiment analyzer required 6 phases: storage of the database of tweets manually labeled; pre-processing, using text cleaning methods, tokenization and stopwords removal; application of the *Term Frequency–Inverse Document Frequency* (TF-IDF) method; classifier training; collecting new tweets in streaming and; real-time sentiment analysis. After the development of the aforementioned phases, the sentiment analyzer reached an accuracy of 76.8 percent.

KEYWORDS: Machine Learning. Natural language processing. Twitter.

INTRODUÇÃO

O mercado de ações tem se mostrado cada vez mais atrativo para as pessoas físicas. O assunto tem extrapolado o âmbito empresarial e de grandes investidores já consolidados, tendo despertado o interesse de pessoas comuns. Tal fato fica evidente ao observar os números publicados por Purchio (2020). O autor informa que, ao final de 2019, havia 1,67 milhão de investidores registrados, ao passo que, já em fevereiro de 2020, Neira e Filgueiras (2020) divulgaram que cerca de 2,24 milhões de pessoas físicas haviam sido registradas como investidores na B3, batendo recorde de registros e a maior alta de todos os tempos.

Em uma análise detalhada, Cutler, Poterba e Summers (1989) puderam perceber que as variações manifestadas nos preços de ações e oscilações nos valores fundamentais de ativos se devem a uma série de fatores, mas que um deles são as notícias que ocorrem no âmbito financeiro e econômico. Em vista disso, é possível observar que os veículos de comunicação especializados no mercado de ações, tanto brasileiro como mundial, têm ganhado grande importância nos últimos anos devido ao público atraído por essa onda crescente de investidores.

Sob esta concepção, realizar uma análise com relação ao sentimento que cada texto de notícia passa para o leitor, cerne do Processamento de Linguagem Natural (PLN), e ciente do impacto causado no mercado de ações, esta análise se torna relevante e se faz como objetivo geral deste trabalho. Partindo de uma base de tweets previamente classificada, esta pesquisa possui 3 objetivos principais.

Primeiramente, o treinamento de um classificador a partir de uma base de *tweets* previamente classificada. As duas próximas, capturar os *tweets* em tempo real utilizando um mecanismo de coleta em *streaming* e realizar a análise de sentimento dos novos *tweets* capturados utilizando uma plataforma de processamento de dados em tempo real.

Além desta introdução, este artigo está estruturado da seguinte maneira: Na seção Material e Métodos, serão detalhados os métodos e tecnologias utilizados para a treinamento classificador, mecanismo de captura e a construção do analisador de sentimento. Na sequência, seção Resultados e Discussões, a acurácia e performance do classificador Naïve Bayes serão apresentados, bem como a discussão sobre os dados apontados. Por fim, as considerações finais deste trabalho serão apresentadas na seção Conclusão.

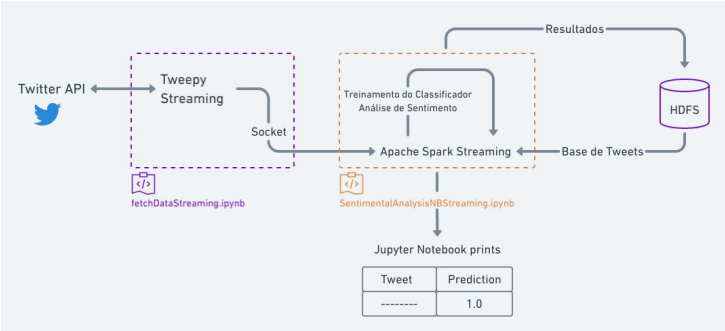
MATERIAL E MÉTODOS

Nesta seção serão descritos e discutidos o material e métodos utilizados neste trabalho. Para sua execução, foram necessárias o uso de algumas tecnologias, que foram: *Python* como linguagem de programação; *Hadoop Distributed File System (HDFS)*, para armazenamento dos *tweets*; *API* do *Twitter juntamente com a biblioteca Tweepy*, manipulando toda a parte de autenticação

e conexão, para realizar a coleta dos *tweets* em *streaming*; *Apache Spark Streaming*, sistema escalonável de processamento de dados, ligadamente com os componentes da API *Spark* (*Mlib*, *SparkSQL e GraphX*), utilizado como ambiente de processamento e manipulação dos *tweets* recebidos e treinamento do classificador Naive Bayes (Zhang, 2004); biblioteca *Socket* para realizar a integração entre *Tweepy* e o *Spark Streaming*.

Este trabalho de pesquisa requereu um fluxo de desenvolvimento em que diversas tecnologias foram necessárias, para isto, a Figura 1 ilustra a integração entre elas. Seus aspectos operacionais serão especificadas subsequentemente.

Figura 1 – Integração entre as tecnologias



Fonte: Autoria Própria (2020)

Primeiramente foi realizado o armazenamento de uma base de tweets, manualmente rotulados pelo trabalho de pesquisa realizado por Melo e Meinerz (2019), somando 2830 tweets coletados no períodos de 12/11/2015 à 16/05/2019, e rotulados quanto a sua polaridade, em um diretório criado dentro do HDFS.

Em seguida, esta base de tweets é submetida a uma fase de pré-processamento, constituída por 4 fases: alocação em um dataframe; limpeza do texto; tokenização e; remoção de stopwords.

A alocação dos tweets é feita em um *dataframe*, tabela que armazenam uma base de dados, no qual cada linha corresponde a um registro (*tweet*) e cada coluna às propriedades a serem armazenadas (Torgo, 2003). Em seguida, são realizados o processo de limpeza do texto, no qual é retirado links e pontuações, e o de tokenização, processo de divisão de um texto em palavras (*tokens*). Por fim, são retiradas as *stopwords*, palavra que aparecem com frequência e não possuem tanto significado.

O próximo passo é realizar o treinamento do classificador Naive Bayes, e para isto, foram necessário o uso de dois métodos. *Term Frequency - Inverse Document Frequency* (TF-IDF) é o primeiro método utilizado, uma técnica utilizada na recuperação de informações e mineração de texto, sendo uma medida estatística que avalia a importância de uma palavra para um documento em um texto ou um corpus, sendo assim, quanto mais vezes uma palavra aparece em um documento, maior a sua importância (Damien, 2016). O primeiro passo é calcular o *Term Frequency* (TF), calculado pela equação (1), e posteriormente, calcular o *Inverse Document Frequency* (IDF), obtido pela equação (2), para então se obter o resultado do método TF-IDF, a partir do produto da equação (1) pela equação (2).

$$TF(t,d) = \frac{Número\ de\ vezes\ que\ o\ termo(t)\ aparece\ no\ documento(d)}{Número\ total\ de\ termos\ no\ documento(d)} \quad (1)$$

$$IDF(t,D) = \log\left(\frac{Número\ total\ de\ documentos(D)}{Número\ de\ documentos\ com\ o\ termo(t)\ nele}\right) \quad (2)$$

O pipeline foi o método aplicado para unir os estágios de pré-processamento realizados anteriormente, unindo-os em um único dataframe, utilizado apenas no período de execução do algoritmo.

Por fim, é aplicado o método Naive Bayes multinomial, algoritmo simples de classificação com um modelo de evento multinomial, seguindo pela equação (3), na qual é aplicado o teorema de Bayes para calcular a distribuição de probabilidade condicional do rótulo e usá-lo para previsão.

$$P(x|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i P_{ki}^{x_i} \quad (3)$$

A biblioteca *Tweepy* foi utilizada como mecanismo de coleta dos novos *tweets*, utilizando métodos de busca e captura em *streaming*, sendo submetido a um processo de filtragem, via expressão regular, para capturar somente *tweets* com os assuntos relacionados as empresas previamente definidas. *StreamListener* é o método responsável por enviar os *tweets via Socket* para o *Spark Streaming*, no qual é submetido a mais um filtro, definindo os veículos de comunicação especializados no mercado de ações brasileiro por meio de seus *Twitter ID*.

Com o modelo gerado a partir do treinamento do classificador, chegou a hora de aplicá-lo aos novos *tweets* coletados. Para tal, os novos *tweets* são recebidos pelo *StreamingContext* do *Spark via socketTextStream*, e enviados para a função responsável por: inseri-los em um *dataframe*, passá-los para o *pipeline*, no qual são executadas as funções de pré-processamento e; por fim, aplica-se a análise de sentimento sobre os *tweets*, utilizando o modelo de Naive Bayes previamente treinado.

Para o armazenamento dos novos *tweets* coletados, foi gerado um novo diretório no *HDFS*, no qual foram alocados todos os objetos dos *tweets* capturados, enviados pela *API* do *Twitter* em formato *JavaScript Object Notation* (JSON).

A visualização dos sentimentos extraídos foi por meio do *Jupyter Notebook*, interface gráfica de desenvolvimento, que permite a utilização do kernel *Python3* para compilação e visualização dos códigos gerados. Nele é possível observar a entrada dos *tweets* no *notebook* em que é configurado os métodos do mecanismo de busca, e a impressão dos resultados da analise de sentimento em tempo real, *notebook* no qual são configurados o *Spark Streaming* e o classificador Naive Bayes.

RESULTADOS E DISCUSSÃO

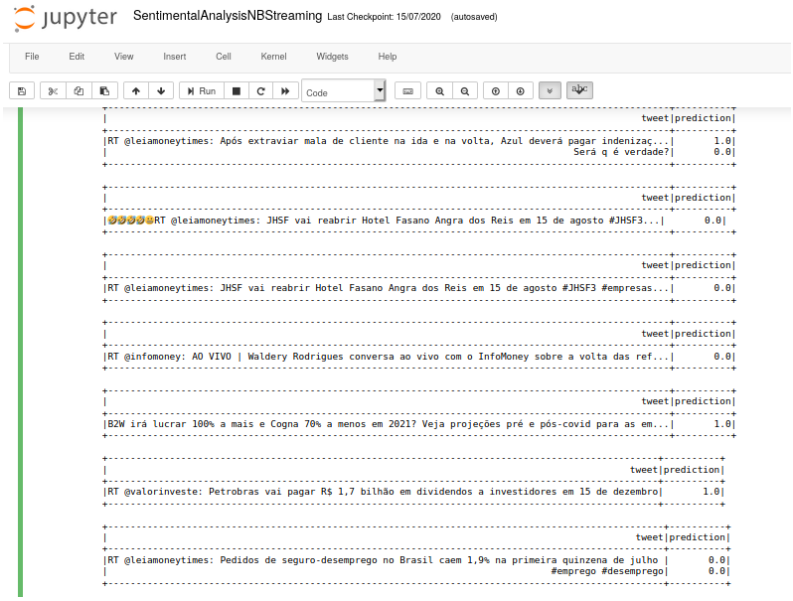
Utilizando da base de tweets previamente classificada, foram aplicados os métodos de pré-processamento e treinamento do classificador, obtendo-se ao final uma coluna no dataframe denominada como "*predictions*", com os resultados, produzindo uma classificação de positivo (índice 1), no qual os textos

presentes nas notícias expressam reações favoráveis a uma determinada empresa, e negativo (índice 0), onde a reação expressa não se demonstra favorável. A precisão do classificador foi medida a partir do método de Multiclass Classification Evaluator, obtendo-se o valor d e 0.768 e variação de ±0.01.

Em razão dos filtros aplicados, o fluxo de tweets capturados pelo mecanismo de busca é baixo. Para alcançar uma acurácia maior do classificador, seria necessário um período maior de captura de novos tweets e classificação, subindo a aplicação para nível de servidor.

Na Figura 2 é possível constatar o resultado analise de sentimento, na qual são impressos os tweets capturados em streaming e o resultado predito pela análise em um dataframe.

Figura 2 – Resultado da Analise de Sentimento



Fonte: Autoria Própria (2020)

CONCLUSÃO

A principal contribuição deste trabalho foi realizar um analisador de sentimento em tempo real de *tweets*, relacionados ao mercado de ações brasileiro, utilizando métodos de *machine learning*, usufruindo da biblioteca *Machine Learning Library* (Mllib) da *Apache Spark*, para treinar um classificador *Naive Bayes*, a partir de uma base de *tweets* previamente classificada, atingindo um nível de acurácia de 76%.

Os resultados desta pesquisa mostraram que, por mais que hajam limitações quanto a análise de sentimento em língua portuguesa, devido a carência de ferramentas, bibliotecas e projetos voltados a este fim, visto a vasta possibilidade de realizar o mesmo ofício em língua inglesa, é possível proceder com um nível de acurácia satisfatório em relação a assertividade humana de 85%.

Tal desfecho sucedeu-se principalmente devido a vasta base de *tweets*, minuciosamente classificada quanto a sua polaridade, provendo um treino mais

favorável ao classificador Naive Bayes, o qual posteriormente aplica a analise de sentimento em tempo real.

AGRADECIMENTOS

Agradeço ao meu orientador, Prof. Dr. Giovani Volnei Meinerz e a Pró-Reitoria de Pesquisa e Pós-Graduação (PROPPG) que, por meio do Programa Institucional de Voluntariado em Iniciação Científica (PIVIC), me oportunizaram o desenvolvimento do pensamento científico, contribuindo para com a minha formação acadêmica.

REFERÊNCIAS

CUTLER, D. M.; POTERBA, J. M.; SUMMERS, L. H. **What moves stock prices? The Journal of Portfolio Management, Institutional Investor Journals Umbrella**, v. 15, n. 3, p. 4–12, 1989. ISSN 0095-4918. DOI: 10.3905/jpm.1989.409212. Eprint: https://jpm.pm-research.com/content/15/3/4.full.pdf. Disponível em: https://jpm.pm-research.com/content/15/3/4. Acesso em: 26 jul. 2020.

DAMIEN. **TF-IDF**. Set. 2016. Disponível em: https://www.beyondthelines.net/machine-learning/tf-idf/. Acesso em: 14 jul. 2020.

MELO, J. G. S. de; MEINERZ, G. V. **Análise de sentimentos de textos voltados ao mercado de ações**. XXIV Seminário de Iniciação Científica e Tecnológica da UTFPR, 2019, Pato Branco, v. 24, p. 0-0, 2019.

NEIRA, A. C.; FILGUEIRAS, I. **Número de pessoas físicas na B3 tem alta recorde e bate 2,24 milhões em março**. Abr. 2020. Disponível em: https://valorinveste.globo.com/objetivo/hora-de-investir/noticia/2020/04/03/numero-de-pessoas-fisicas-na-b3-tem-alta-recorde-e-bate-224-milhoes-em-marco.ghtml. Acesso em: 25 jul. 2020.

PURCHIO, L. **A era da bolsa**. Jan. 2020. Disponível em: https://istoe.com.br/a-era-da-bolsa. Acesso em: 25 jul. 2020.

TORGO, L. **Data Frames**. Out. 2003. Disponível em: https://www.dcc.fc.up.pt/~ltorgo/SebentaR/HTML/node16.html. Acesso em: 12 jul. 2020.

ZHANG, Harry. **The Optimality of Naive Bayes.** 2004, [S.l.]: AAAI Press, 2004.