https://eventos.utfpr.edu.br//sicite/sicite2019

Análise de sentimento de *tweets* referentes a empresas da B3 utilizando o algoritmo K-NN

Sentiment analysis of B3 companies' tweets using the K-NN algorithm

Vitor Fabrile Guastala
vitorguastala@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

Atualmente há um grande número de dados no planeta, que vem crescendo exponencialmente. Os projetos de Big Data, por meio de tarefas como a análise de sentimento, visam manipular esse grande volume de dados a fim de obter resultados úteis para tomada de decisão. Tais projetos podem ser aplicados em diversas áreas, como o mercado de ações. Nesta pesquisa, 3994 *tweets* relacionados a empresas listadas na bolsa de valores brasileira, a Brasil, Bolsa, Balcão (B3) foram coletados, armazenados, anotados, preparados e utilizados para treinamento do algoritmo de classificação K-NN que, após o aprendizado, se mostrou apto para rotular outros *tweets*. Desse modo, investidores teriam um parâmetro adicional ao analisar empresas que participem da B3. Ao aplicar o K-NN, observou-se que a sua acurácia foi de 75%, resultado esse considerado satisfatório, com potencial de ser ampliado, caso seja aplicado a uma base de dados balanceada, com o mesmo número de *tweets* negativos, positivos e neutros.

PALAVRAS-CHAVE: Aprendizado de máquina. Classificação. Mercado de ações.

Recebido: 19 ago. 2019.
Aprovado: 01 out. 2019.

Direito autoral: Este trabalho está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.

ABSTRACT

Currently there is a large number of data on the planet, which has been growing up exponentially. Big data projects, through tasks such as sentiment analysis, aim to manipulate this large amount of data for useful decision-making results. Such projects can be applied in various areas such as the stock market. In this research, 3994 tweets related to companies listed on the Brazilian stock exchange, Brazil, Bolsa, Balcão (B3) were collected, stored, annotated, prepared and used to train the K-NN classification algorithm that, after learning, became showed fit to label other tweets. Thus, investors would have an additional parameter when looking at companies participating in B3. When applying K-NN, it was observed that its accuracy was 75%, a result considered satisfactory, with potential to be expanded if applied to a balanced database, with the same number of negative, positive and neutral tweets.

KEYWORDS: Machine learning. Classification. Stock market.

INTRODUÇÃO

De acordo com Mendes (2016), cerca de 90% dos dados de todo o planeta foram gerados nos últimos dois anos, e esse volume de dados vem crescendo exponencialmente. Além disso, 80% desses dados são não-estruturados, ou seja, não possuem uma estrutura pré-definida, ou estão em diferentes formatos. Com a necessidade de interpretar e analisar esse grande conjunto de dados, surge o conceito de Big Data, unido à análise de sentimento.

Big Data são metodologias para capturar, armazenar, processar, analisar e visualizar essa grande quantidade de dados, a fim de obter respostas para problemas com rapidez (MONTINI, 2018). Já a análise de sentimento, é a extração de informações sobre comportamento e sentimento a partir de texto. A classificação atribui um valor discreto para cada um dos conteúdos textuais analisados, como positivo, negativo ou neutro (ALVES, 2015).

Um projeto de Big Data pode ser aplicado em diversas áreas, como o mercado de ações. Ser capaz de analisar as ações de forma apropriada, determina o sucesso dos melhores investidores (INFOMONEY, 2005). Por isso, esta pesquisa analisou *tweets* em português brasileiro, referentes à empresas listadas na bolsa de valores brasileira, a Brasil, Bolsa, Balcão (B3), com o objetivo de realizar a análise de sentimento, aplicando um método de aprendizado de máquina supervisionado, com o intuito de classificar *tweets* automaticamente.

Optou-se por analisar *tweets* em português brasileiro pois, segundo Statista (2019), o Brasil é o sexto país com mais usuários na rede social Twitter, com 8,28 milhões de usuários até julho deste ano.

Quanto ao método de aprendizado de máquina supervisionado, o classificador *K-Nearest Neighbors* (K-NN) foi treinado para aplicar a análise de sentimento em outros *tweets*. O K-NN foi escolhido por ser um algoritmo cujo processo de classificação é considerado simples, e também por sua implementação ser considerada elementar, além de sua aplicação no setor financeiro ser bastante comum (JAAFAR; MUKAHAR; RAMLI, 2016; IMANDOUST; BOLANDRAFTAR, 2013).

MATERIAL E MÉTODOS

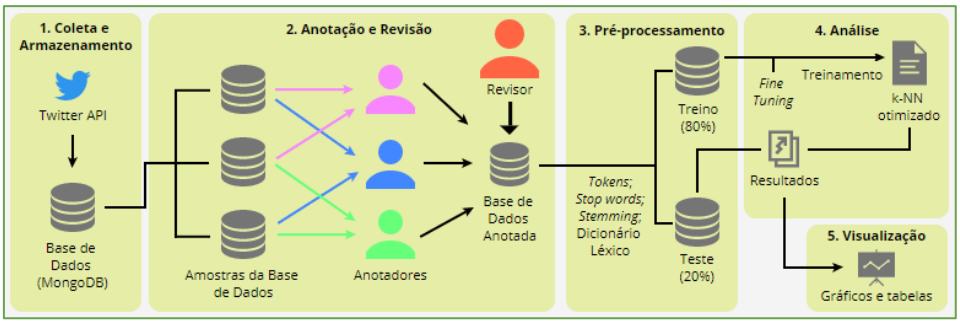
Para a execução desta pesquisa, foram utilizadas as seguintes tecnologias: Twitter API, MongoDB, Python, e bibliotecas NLTK, Scikit-Learn, Plotly e Matplotlib.

Com a API oferecida pelo próprio Twitter, foi possível coletar os *tweets* de interesse e armazená-los no MongoDB, um Sistema de Gerenciamento de Banco de Dados (SGBD) orientado a documentos, que facilita o armazenamento e manipulação de dados não-estruturados, que é o caso dos *tweets*.

Com a linguagem de programação Python, foi possível ter fácil acesso à diversas bibliotecas úteis relacionadas a Inteligência Artificial, sendo elas a NLKT para o processamento de dados, a Scikit-Learn para a análise e Plotly e Matplotlib para a visualização.

A Figura 1 mostra um esquema em que se resume todo o método aplicado neste estudo, desde a coleta de dados, até a visualização dos resultados.

Figura 1 – Esquema de todo método utilizado na pesquisa



Fonte: Autoria Própria (2019).

Um *crawler* em Python foi desenvolvido para se conectar com a API do Twitter, e então foram **coletados** *tweets* em português brasileiro, que mencionassem as empresas listadas na B3, publicados entre 16 de Novembro de 2015 e 16 de Maio de 2019, por veículos de comunicação especializados no domínio de mercado de ações (Estadão Economia, Valor RI, etc.).

Uma ***string* de busca** foi criada para filtrar os nomes ou códigos das empresas na B3, sendo que *tweets* que continham mais de uma empresa citada foram descartados para reduzir a complexidade de anotação dos dados. No total, foram armazenados 3994 registros no MongoDB, cada registro possuindo: um identificador, data de criação, empresa a qual o *tweet* se refere, texto/descrição do *tweet*.

Para obter o **sentimento** a que o *tweet* se refere, foi necessário submetê-los a um processo de **anotação**, no qual três alunos e um professor orientador, especialista no mercado de ações, assumiram o cargo de anotadores. A fim de facilitar o processo de anotação, tornando-o mais rápido, dividiu-se a base de dados em três amostras de tamanhos iguais, melhor compreendido na Figura 1.

Cada *tweet* foi lido e interpretado por cada anotador, que definia o seu sentimento. Os anotadores podiam também classificá-los como confusos, caso não conseguissem definir uma polaridade, ou irrelevantes, caso o *tweet* não devesse ter sido coletado devido à uma falha na *string* de busca, por exemplo. Apenas *tweets* em que havia acordo entre os anotadores foram mantidos. *Tweets* confusos e irrelevantes foram descartados. Com isso, o número total de registros foi reduzido para 2831, o que diminuiu cerca de 30% da base de dados.

Em seguida, os *tweets* passaram por um processo de revisão, em que o professor especialista ficou responsável por analisar os *tweets* a fim de verificar se a polaridade definida pelos alunos estava correta.

Utilizando a biblioteca NLKT, a próxima etapa foi a de **pré-processamento** dos dados, que consiste em transformar os *tweets* em informações úteis e de fácil interpretação para um computador. Foram realizadas 5 fases:

a) ***Tokens***: foram utilizadas expressões regulares (regex) para substituir padrões nos *tweets*, como URLs, *hashtags*, datas, números, nomes de usuários e de empresas;

b) **Lista de palavras**: os *tweets* foram transformados em listas de palavras, removendo a pontuação;

c) ***Stop words***: removeu palavras sem significado relevante para a análise;

d) ***Stemming***: processo em que se substitui cada palavra por sua raiz, por exemplo: “amável”, “amor” e “amoroso” tornam-se “am”;

e) **Dicionário léxico**: criado para substituir termos que possivelmente confundiriam o algoritmo de classificação, tornando termos complexos em termos simples.

Todo esse processo diminuiu a quantidade de palavras, e consequentemente a quantidade de características, por isso sua importância. Sem a preparação dos dados, haviam 6739 características e, após a preparação, o número caiu para 2296.

Após o término do pré-processamento, foi criada a *bag of words*, em que todas as palavras únicas de todos os *tweets* processados foram transformadas em um vetor de características, denominado **descritor**. Com isso, para cada *tweet*, caso este possuísse determinada palavra, teve valor 1 para aquela palavra, do contrário, teve valor 0, tornando o *tweet* um objeto de 2296 dimensões, ou seja, um *array* com 2296 posições, em que cada posição representa uma característica.

Foi realizada a validação por *holdout*, em que 80% da base de dados foi utilizada como treino para o classificador K-NN, ou seja, utilizada para ensinar o classificador acerca de como rotular os *tweets* corretamente, enquanto os outros 20% foram utilizados para teste, a fim de verificar se o classificador aprendeu a rotular outros *tweets*.

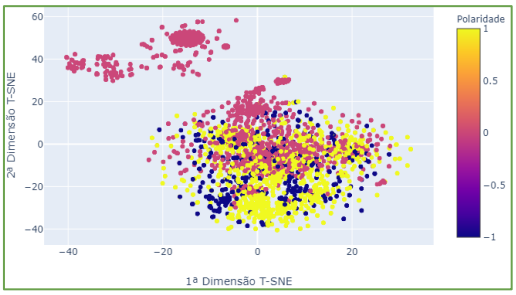
A próxima etapa executada foi a de **análise**, empregando o classificador K-NN, disponível na biblioteca Scikit-Learn. Para que haja a escolha dos melhores parâmetros a serem passados para o classificador K-NN, visto que estes variam de acordo com o domínio de negócio e são empíricos, foi realizado o ***fine tuning*** (ajuste fino). Trata-se de um método em que são testadas diversas combinações entre possíveis parâmetros, tendo como retorno os parâmetros que obtiveram melhor acurácia na classificação.

RESULTADOS E DISCUSSÃO

O principal parâmetro do K-NN é o valor de k, que corresponde ao número de vizinhos próximos que o algoritmo deve considerar para a classificação, e o melhor valor definido pelo *fine tuning* foi 4, com 0.7298 de acurácia no treino.

Utilizando uma técnica de redução de dimensionalidade (T-SNE), é possível ter uma visualização da polaridade dos *tweets* da base de dados, como mostra a Figura 2, sendo, em amarelo, *tweets* positivos (989), em azul, negativos (396) e em rosa, neutros (1445).

Figura 2 – Visualização da polaridade dos *tweets* utilizando T-SNE



Fonte: Autoria Própria (2019).

A Tabela 1 mostra os resultados obtidos no conjunto de teste após o treinamento do algoritmo K-NN com o melhor conjunto de parâmetros definido pelo *fine tuning*. Foram realizados três tipos de teste, que se diferenciam quanto a preparação dos dados.

Tabela 1 – Resultados obtidos com K-NN treinado

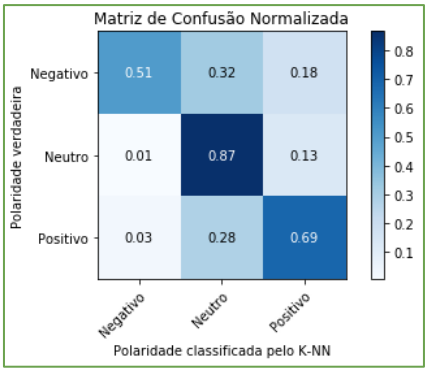
Tipo de teste	Acurácia	Precisão	*Recall*	F1	Tempo (s)
Não utilizando dicionário	0.7473	0.7553	0.7473	0.7377	25.94
Não utilizando *tokens*	0.7402	0.7401	0.7403	0.7321	229.73
Utilizando ambos	**0.7527**	**0.7564**	**0.7527**	**0.7468**	**20.13**

Fonte: Autoria Própria (2019).

Esses resultados mostram a importância dos *tokens* e do dicionário na preparação dos dados. Pode-se dizer que os valores obtidos são satisfatórios, principalmente com relação ao parâmetro F1, que leva em conta que a base de dados não é balanceada, uma vez que houveram mais *tweets* positivos coletados do que negativos (Figura 2), o que interfere principalmente na acurácia.

Isso pode ser melhor visualizado na Figura 3, que mostra a matriz de confusão no conjunto de teste.

Figura 3 – Matriz de confusão no conjunto de teste



Fonte: Autoria Própria (2019).

A matriz de confusão deve ser interpretada de modo que as colunas representam a polaridade que o classificador determinou para aquele *tweet*, enquanto que as linhas representam a polaridade verdadeira daquele *tweet*.

Dessa forma, possivelmente por conta do desbalanceamento da base de dados, observa-se que houve um erro maior com relação aos *tweets* negativos, mas mesmo assim o classificador obteve uma boa porcentagem de acurácia, como visto na Tabela 1.

CONCLUSÃO

Analisar o mercado de ações é uma tarefa que todo investidor precisa realizar, e ter um mecanismo que possa auxiliar na tomada de decisões é de suma importância. Como visto na seção resultados, o K-NN trouxe uma boa porcentagem

de acurácia com relação ao teste, mesmo apresentando alguns erros de acordo com a matriz de confusão.

Apesar de erros serem sempre esperados, pode-se melhorar essa acurácia coletando mais *tweets*, a fim de balancear melhor a base de dados, obtendo o mesmo número de *tweets* negativos, positivos e neutros, ou ainda, aprofundando a pesquisa com relação ao K-NN, procurando uma maneira do algoritmo considerar o desbalanceamento do conjunto de treino.

Pode-se concluir que este estudo teve um resultado satisfatório na análise do sentimento embutido em *tweets*, identificado por meio da aplicação do algoritmo K-NN, uma vez que há poucas pesquisas relacionadas ao domínio do mercado de ações de empresas listadas na B3.

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**. Brasília: UnB, 2015.

IMANDOUST, S. B.; BOLANDRAFTAR, M. **Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background**. International Journal of Engineering Research and Applications, [S.l.], v. 3, p. 605–610, 2013.

INFOMONEY. **Como analisar o mercado de ações**. InfoMoney, [S.l.], 31 dez. 2005. Disponível em: <https://www.infomoney.com.br/educacao/guias/noticia/527579/como-analisar-mercado-acoes>. Acesso em: 29 jul. 2019.

JAAFAR, H. b.; MUKAHAR, N. b.; RAMLI, D. A. b**. A methodology of nearest neighbor: Design and comparison of biometric image database**. *In*: 2016 IEEE Student Conference on Research and Development (SCOReD). [S.l.: s.n.], 2016. p. 1–6.

MENDES, D. **Big Data Fundamentos 2.0**. Data Science Academy, Brasília, 2016. Disponível em: <https://www.datascienceacademy.com.br/course?courseid=big-data-fundamentos>. Acesso em: 29 jul 2019.

MONTINI, A. **Introdução ao Big Data**. Fundação Instituto de Administração, São Paulo, 2018. Disponível em: <https://www.coursera.org/lecture/introducao-big-data/bloco-1-definicao-do-big-data-Gld6r>. Acesso em: 29 jul 2019.

STATISTA. **Leading countries based on number of Twitter users as of July 2019 (in millions)**. Statista, Hamburgo, 2019. Disponível em: <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>. Acesso em: 29 jul. 2019.