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

Análise de sentimentos de *tweets* do mercado de ações brasileiro utilizando *Naive Bayes*.

Sentiment analysis of tweets about brazilian stock market using Naive Bayes.

Thaysla Fernanda Gomes da Cruz
thayslacruz@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasill

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

RESUMO

Acionistas da Brasil, Bolsa, Balcão (B3) se preocupam em estudar e analisar o contexto das empresas cotadas na B3 a fim de decidir o melhor momento de investir, por isso, veículos de informação especializados no domínio do mercado de ações divulgam manchetes de suas notícias através da rede social *Twitter*. Esse artigo descreve a aplicação do algoritmo *Naive Bayes* (NB) para desempenhar uma análise de sentimentos desses *tweets* e classificá-los quanto a sua polaridade positiva, negativa ou neutra. Para realização do processamento de linguagem natural e classificação textual foram realizadas 5 etapas: coleta, armazenamento, anotação, pré-processamento e análise dos dados, ademais, na fase de pré-processamento foram empregadas expressões regulares e dicionário léxico, desenvolvidos especificamente para o domínio do mercado de ações brasileiro. Com isso, o classificador de aprendizado de máquina supervisionado NB atingiu uma acuŕacia de 78 por cento, poupando a necessidade dos acionistas lidarem com um grande volume de dados ante a possibilidade de classificação automatizada.

PALAVRAS-CHAVE: Aprendizado do computador. Big data. Inteligência artificial. Processamento de linguagem natural.

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

ABSTRACT

Investors of Brasil, Bolsa, Balcão (B3) care about study the companies in B3 list to decide the best moment to invest, therefore, information vehicles share your headlines on Twitter. This article describe the algorithm Naive Bayes (NB) application to analyze the tweets content and classify it as positive, negative or neutral. For a natural language processing realization was necessary 5 phases: data collect, storage, annotation, preprocessing and analysis, in addition, was applied regular expressions and lexicon dictionary specifically created to brazilian trading market on preprocessing phase. The supervised classifier NB reached 78 percent of accuracy, sparing the investors the need of deal with a huge data in view of the possibility of an automatic classification.

KEYWORDS: Machine learning. Big data. Artificial intelligence. Natural language processing.

INTRODUÇÃO

As bolsas de valores são responsáveis por gerir a negociação de ações das empresas de capital aberto. Tais empresas disponibilizam títulos de propriedade de parte de seu capital social para serem negociados. No Brasil, atualmente, esse mercado é representado pela empresa Brasil, Bolsa, Balcão (B3), a bolsa de valores brasileira (KONKERO, 2018).

A B3 possui índices que atuam como indicadores de desempenho, porém, para operar na bolsa de valores deve-se realizar uma análise que vai além desses índices, pois o cenário está em constante mudança. O mercado de ações é fundamental para a economia de um país, portanto, há muitas notícias sendo divulgadas. Um aglomerado dessas, pode ser encontrado no *Twitter*, uma vez que o Brasil é o segundo país que mais possui usuários nessa rede social (ASLAM, 2019).

Ao lançar uma nova publicação, sua manchete e *link* para acesso são divulgados através de um *tweet* e a leitura desses oferece um contexto confiável. Designar esse trabalho para o algoritmo *Naive Bayes*, torna o processo mais rápido e menos custoso. Ele possui a capacidade de realizar uma análise e classificar *tweets* dentre as polaridades positiva, negativa ou neutra. Evitando que uma pessoa se submeta a leitura de um grande volume de dados e o risco de cometer erros.

MATERIAS E MÉTODOS

O algoritmo de aprendizado supervisionado ***Naive Bayes*** é um classificador probabilístico. Assim como o teorema de Bayes, assume que características são completamente independentes, então são aprendidas separadamente (PARVEEN; PANDEY,2016).

O algoritmo classifica os dados em duas etapas. Na primeira etapa, de treinamento, o método faz uso das amostras fornecidas para aprendizado e seleciona as características de cada classe. Na segunda etapa, de predição, para cada novo dado o método calcula a hipótese de cada classe. A classe designada será a que possui a maior probabilidade (IBRAHIM; YUSOFF, 2019).

Para implementação do algortimo foi realizada a **coleta de dados**, através da **API do *Twitter*,** associada a um ***crawler*** na **linguagem de programação *Python*.** Foram estabelecidas *strings* de busca que restringem a coleta apenas a *tweets* que fazem menção às empresas cotadas na B3, sendo descartados aqueles nos quais constam mais de uma delas.

A primeira etapa nos proporcionou 3994 *tweets*, postados na rede social em um intervalo compreendido entre o dia 16 de Novembro de 2015 até 16 de Maio de 2019, por diferentes fontes especializadas no domínio do mercado de ações, sendo algumas delas: Estadão Economia e *Money Times*. Todos os *tweets* coletados foram **armazenados** em um banco de dados de um SGBD orientado a documentos, denominado **MongoDB**.

Foi necessário realizar um processo de **anotação manual** para efetuar a fase de treinamento, cada um dos 3994 *tweets* foram rotulados de acordo com a sua polaridade positiva, negativa ou neutra. Esse processo foi realizado por 3 alunos, em que cada aluno rotulou 2 terços das amostras, sendo descartados *tweets* em que não houve concordância quanto a polaridade e, como consequência, restaram 2830 deles.

Por fim, foi realizada uma revisão de todo o *dataset* por um quarto anotador com experiência no supracitado domínio. Após a revisão houve a percepção de que a coleta não proporcionou a mesma quantidade de *tweets* de cada polaridade, portanto a amostra está desbalanceada.

Após a revisão foi iniciada a etapa de **pré-processamento**, por meio da **biblioteca NLTK.** O primeiro passo foi obter cada palavra separadamente através da *tokenization*. Assim, foi possível a remoção de *stop words*, ou seja, palavras frequentes que não possuem relevância e posteriormente a redução das palavras restantes ao seu radical, por meio do *stemming*.

Foram aplicadas **expressões regulares** (regex), para que textos sem carga semântica fossem desconsiderados. Isto posto, numerais, símbolos, datas, URLs, *hashtags* e fotos foram substituídos por um *token*. Em seguida, foi aplicado o **dicionário léxico**, composto de termos do domínio financeiro, que ao se encontrarem nesse contexto assumem sentido diferente.

Os *tweets* foram convertidos a vetores do tamanho da lista de características, apresentam o valor 1 se a característica está presente e 0 se não. Sendo a lista de características todas as palavras únicas de todo o conjunto de treinamento. Por meio da fórmula do teorema de Bayes, ilustrada na Eq. (1) é obtida a probabilidade do *tweet* pertencer classe A, uma vez que a característica B está presente, assumindo todas as características como independentes (TAHERI; MAMMADOV, 2013).

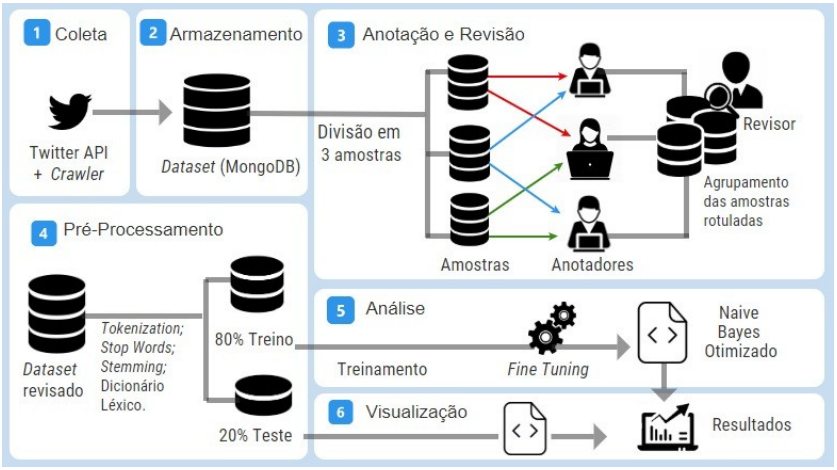
$$P(A \vee B) = \frac{P(B \vee A)\,P(A)}{P(B)} \qquad (1)$$

A biblioteca **Scikit-Learn** propiciou a implementação do algoritmo ***Complement Naive Bayes* (CNB)**, que tem como critério a fequência que as características aparecem, porém, considera o desbalanceamento da amostra (YONGCHENG, 2018). O algoritmo utiliza parâmetros para melhorar sua classificação, foi realizado o *Fine Tuning* (ajuste fino), que por meio de diversos testes definiu uma solução. Por fim, foi gerada a **visualização** dos resultados obtidos através da classificação do algoritmo.

O fluxograma da Figura 1 representa todas as etapas desenvolvidas neste projeto. Desde a coleta, até a divisão do *dataset* em 3 amostras, para que cada auno realizasse a rotulação manual de 2 delas. Após isso, as amostras foram agrupadas e revisadas, gerando o *dataset* final, que após o pré-processamento foi dividido entre 80% para treino e 20% para teste. Na fase de treinamento o algoritmo foi otimizado através

do *Fine Tuning*, e por fim, os testes foram realizados e obteve-se a visualização dos resultados.

Figura 1 – Etapas do projeto

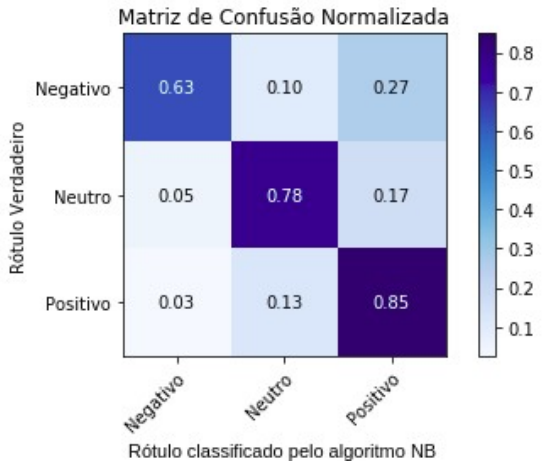


Fonte: Autoria própria (2019).

RESULTADOS E DISCUSSÃO

Ao classificador CNB foram fornecidos 2264 *tweets* para fase de treinamento e 566 para teste, sendo 35% positivos, 14% negativos e 51% neutros. Com isso, o algoritmo obteve uma acurácia de 78,2%. Através da matriz de confusão representada na Figura 2, pode-se identificar quais são os pontos fracos e fortes do algoritmo. Seu menor desempenho está em reconhecer dados negativos, com 63% de acerto, pois é a classe representada em menor quantidade no *dataset.*

Figura 1 – Matriz de confusão



Fonte: Autoria própria (2019).

O conjunto de teste é treinado com as características obtidas no conjunto de treinamento, portanto, há uma perda de informações quanto as palavras exclusivas do conjunto de testes.

Por se tratar de uma amostra relativamente pequena, essa perda ocorreu em maior escala na classe dos negativos.

Além disso, muitos dos erros apresentados pelo algoritmo são devidos à *tweets* que apresentam palavras comumente positivas em casos negativos, ou vice-versa. Como exemplo, há um tweet que afirma "Vale tem leve alta após registrar prejuízo bilionário". Apesar de uma pessoa identificar facilmente que é uma notícia positiva, para o CNB a presença da característica "prejuízo" induz ao erro, classificando-a como negativa.

Para exibir a contribuição do dicionário léxico e das regex no processo de clasificação, a aplicação do algoritmo foi realizada sem o uso de cada uma delas e, por fim, sem ambas as técnicas. conforme pode ser observado na Tabela 1.

Sem o uso de regex houve uma perda de desempenho, que pode ser notada através do tempo de classificação, que foi maior que o dobro. Isso ocorre devido a quantidade de características que foi de 2296 para 5340. Removendo o dicionário léxico, diferente da regex, a acurácia diminui pois o algoritmo deixa de considerar o valor real que determinadas palavras assumem ao serem aplicadas ao mercado financeiro.

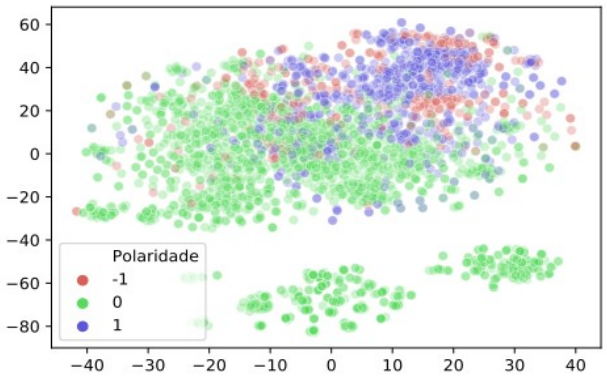
Tabela 1 – Aplicação de diferentes técnicas

Modelo	Acurácia	Precisão	Recall	F1 score	Tempo
Completo	78,26%	79,19%	78,26%	78,33%	0.0043 s
Sem regex	78,09%	78,61%	78,09%	78,10%	0.0092 s
Sem dicionário	77,56%	78,84%	77,56%	77,60%	0.0057 s
Sem regex/ dicionário	77,20%	77,89%	77,20%	77,18%	0.0088 s

Fonte: Autoria Própria(2019).

A Figura 3 nos permite ver o quanto os *tweets* com mesma polaridade são semelhantes pois se apresentam agrupados, provando o quão aplicável são os classificadores nesse domínio. Em contrapartida, pode-se observar que os negativos estão dispersos, denotando a dificuldade do algoritmo.

Figura 3 – Regiões de polaridade



Fonte: Autoria própria (2019).

CONCLUSÃO

Investidores da B3 dedicam horas à análise das empresas a fim de compreender o momento certo de comprar e vender suas ações. Nessa pesquisa o algoritmo *Naive Bayes* foi utilizado para classificar *tweets* com notícias referentes a essas empresas quanto a sua polaridade positiva, negativa ou neutra.

Através da classificação obteve-se uma assertividade de 78,26% com resultados instantâneos, assertividade semelhante a capacidade humana que se apresenta entre 72% e 85% (NASCIMENTO et al., 2012). Poupando o tempo dos investidores e a necessidade de lidar com um grande volume de dados.

Uma vez que o algoritmo já está treinado, em trabalhos futuros pode ser utilizado para tratar cada setor ou empresa separadamente, para que sejam realizadas análises preditivas quanto a eles.

REFERÊNCIAS

ASLAM, Salman. **Twitter by the Numbers: Stats, Demographics & Fun Facts.** 6 jan. 2019. Disponível em: https://www.omnicoreagency.com/twitter-statistics/. Acesso em: 13 ago. 2019.

IBRAHIM, Mohd Naim Mohd; YUSOFF, Mohd Zaliman Mohd. **Twitter sentiment classification using Naive Bayes based on trainer perception**, 2015 IEEE Conference on e-Learning, e-Management and e-Services (IC3e), Melaka, 2015, pp. 187-189.

KONKERO. **Bolsa de Valores / Bovespa - o que é e qual o seu significado**. 25 set.2018. Disponível em:https://www.konkero.com.br/financas-pessoais/economizar/bolsa-de-valores-bovespa-o-que-e-e-qual-o-seu-significado. Acesso em:1 ago 2019.

NASCIMENTO, Paula et al. Análise de sentimento de tweets com foco em notícias. **Análise de Sentimento de Tweets Com Foco em Notícias.** São Paulo, p. 600-613. jul.2012.

TAHERI, Sona; MAMMADOV, Musa. Learning the naive Bayes classifier with optimization models. **International Journal of Applied Mathematics and Computer Science**, 2013, v. 23, n. 4, pp. 787-795.

YONGCHENG, Wu. **A New Instance-weighting Naive Bayes Text Classifiers**, 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), China, 2018, pp. 198-202.