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

Análise de sentimentos de textos voltados ao mercado de ações

Sentiment analysis of stock market texts

João Guilherme Squinelato de Melo
joaomelo@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 emprego do algoritmo SVM para classificar o humor de textos oriundos do microblog Twitter. Como objeto de estudo, tem-se *tweets* de notícias publicadas por veículos de comunicação especializados no domínio do mercado de ações brasileiro. Para o desenvolvimento do classificador foram executadas 6 fases: coleta; armazenamento; anotação; pré-processamento; análise e; visualização. Além disso, duas etapas pertencentes a fase de pré-processamento foram personalizadas para a realização da Análise de Sentimentos. Essas etapas se referem a busca e substituição de padrões no texto e quanto ao dicionário léxico, etapas adaptadas ao contexto do mercado de ações brasileiro. Após o desenvolvimento e execução das fases supracitadas, o classificador atingiu uma acurácia de 83,2155 por cento, o que representa uma assertividade dentro da atingida por humanos: 72 a 85 por cento.

PALAVRAS-CHAVE: Inteligência artificial. Processamento de linguagem natural. Twitter.

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

ABSTRACT

THIS ARTICLE DESCRIBES THE USE OF THE SVM CLASSIFIER TO RATE THE MOOD OF TWITTER MICROBLOG TEXTS OF BRAZILIAN STOCK MARKET. THIS WAS ACHIEVED THROUGH 6 PHASES: COLLECTION; STORAGE; ANNOTATION; PREPROCESSING; ANALYSIS AND; VISUALIZATION. FURTHERMORE, TWO TECHNIQUES WERE PROPOSED: THE SEARCH AND SUBSTITUTION OF PATTERNS IN THE TEXT AND A LEXICAL DICTIONARY. AS A RESULT, THE CLASSIFIER REACHED AN ACCURACY OF 83.2155 PERCENT, WHICH REPRESENTS AN ACCURACY WITHIN THAT ACHIEVED BY HUMANS: 72 TO 85 PERCENT.

KEYWORDS: Artificial Intelligence. Natural language processing. Twitter.

INTRODUÇÃO

Desde a ascensão da Internet a quantidade de dados gerados digitalmente todos os dias tomou proporções gigantes, ao ponto de que destes, 90% foram criados só nos últimos dois anos. Em contrapartida, de toda essa quantidade de dados, apenas uma pequena quantia torna-se objeto de análises. Sob um prisma otimista, esse fato representa uma série de oportunidades de pesquisa e análise sobre esses dados (MARQUESONE, 2016).

Dentre essas oportunidades, existe a Análise de Sentimentos. Ramo da Inteligência Artificial (IA), dentro do subgrupo de Processamento de Linguagem Natural (PLN) interessada em descobrir o sentimento embutido em um texto. Sob este prisma, Bollen, Mao e Zeng (2011) utilizaram da Análise de Sentimentos de *tweets*, postagens do microblog Twitter, para auxiliar no processo de predição das oscilações do mercado de ações.

À vista disso, este trabalho tem por objetivo realizar a Análise de Sentimento de *tweets* envolvendo o mercado de ações brasileiro. Um dos motivos se dá devido ao Brasil ser o sexto país em número de usuários (STATISTA, 2019) do Twitter. Outra justificativa está no do idioma português ser o quinto mais falado no Twitter (STATISTA, 2013), o que suprime a criação de uma base de dados de *tweets* envolvendo o mercado de ações brasileiro.

Para a realização da Análise de Sentimentos, optou-se pelo classificador *Support Vector Machine* (SVM), para classificar *tweets* quanto a sua polaridade, isto é, quanto ao seu humor positivo, negativo ou neutro. A escolha se baseia nos bons resultados atingidos pelo classificador SVM para com a Análise de Sentimentos, como mostram Zainuddin e Selamat (2014).

Entretanto, trabalhar com Análise de Sentimentos em português, denota desafio. Isso em razão da carência de bases de dados e ferramentas linguísticas auxiliadoras (KANSAON; BRANDÃO; PINTO, 2018) para português. Logo, com este trabalho, além da implementação classificador SVM, apresenta-se a criação de três contribuições.

A primeira, a criação de uma base dados de *tweets* envolvendo o mercado de ações. As duas outras, a criação de um dicionário léxico e de um mecanismo de busca e subtituição de padrões em *tweets*, personalizados para o mercado de ações brasileiro.

MATERIAL E MÉTODOS

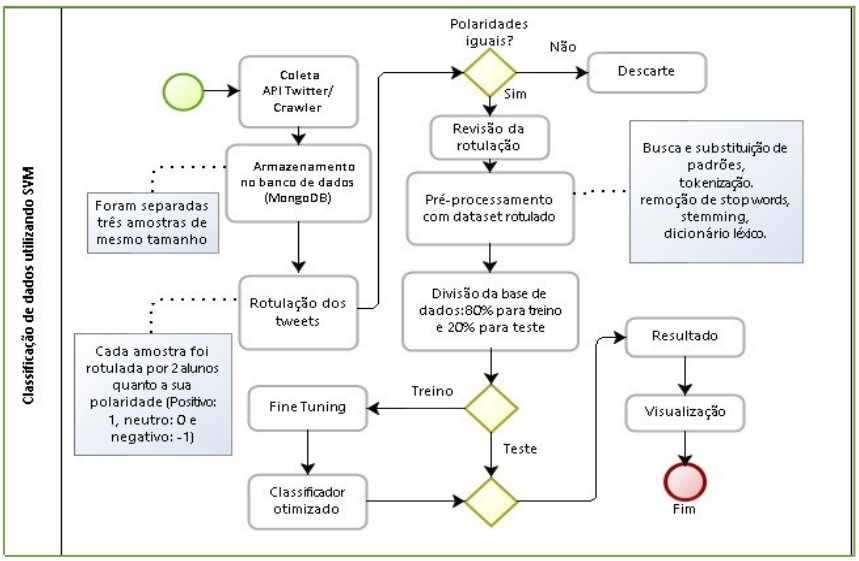
Para a realização deste trabalho de pesquisa, foram utilizadas diversas tecnologias, quais sejam: API do twitter, utilizado para a coleta de *tweets*; MongoDB, responsável por armazenar os *tweets*; Python como linguagem de programação; biblioteca NLTK para realizar o PLN; biblioteca Scikit-learn, para executar a análise dos dados; bibliotecas Plotly e Matplotlib para a visualização dos resultados.

O Sistema de Gerenciamento de Banco de Dados (SGBD) MongoDB foi escolhido devido à sua forma de armazenamento, orientada a documentos. Este SGBD torna fácil a manipulação de dados **não estruturados**, como são os *tweets*.

Já a escolha por Python dá-se em virtude da existência de diversas bibliotecas úteis para a realização das fases de pré-processamento, análise e visualização dos dados. Além disso, com Python, pôde-se criar um *crawler* responsável pela coleta dos *tweets*, o qual foi possível com o uso da API do Twitter.

A Figura 1 ilustra todo o fluxo de desenvolvimento do classificador. Nela, vê-se as fases de coleta, armazenamento, anotação, pré-processamento e análise dos dados.

Figura 1 – Fluxo de Desenvolvimento do Classificador



Fonte: Autoria própria (2019).

Foram coletados *tweets* compreendidos entre 16 de Novembro de 2015 e 16 de Maio de 2019, no idioma **português brasileiro**. Todavia, apenas foram coletados *tweets* publicados por veículos de comunicação especializados no domínio do mercado de ações, tais como Reuters Brasil, Valor RI, dentre outros.

Além disso, apenas os *tweets* que mencionassem empresas listadas na bolsa de valores brasileira (B3) foram coletados. Almejando reduzir a complexidade da análise, *tweets* que mencionassem mais de uma empresa foram descartados.

Em seguida, os *tweets* coletados foram armazenados em um banco de dados do SGBD MongoDB, totalizando **3994 registros** (*tweets*). Entretanto, nenhum tratamento prévio nos registros foi necessário para o armazenamento. Isso, devido a natureza não relacional do referido SGBD, próprio para manipular dados não estruturados de forma nativa.

No entanto, como os *tweets* são coletados sem rótulo, isto é, sem polaridade, é necessários que estes sejam **anotados**. Assim, o processo de anotação atribui aos *tweets* um rótulo que expresse o humor por trás do texto, sendo eles: positivo, negativo ou neutro.

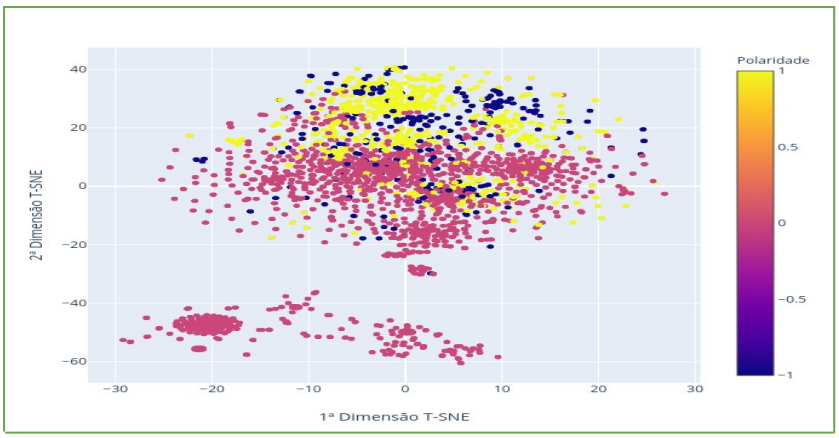
Todavia, o processo de anotação é realizado **manualmente**, porém, com o auxílio de uma planilha eletrônica. Para isso, foram designados três alunos pesquisadores para a anotação, sem experiência no domínio, e um professor orientador para revisão, com experiência no domínio.

Posteriormente, a base de dados de *tweets* foi dividida em três amostras de tamanhos iguais, para diminuir a carga de trabalho. Desse modo, cada aluno se encarregou de anotar dois terços da base de dados de *tweets*, onde cada terço foi anotado por dois alunos.

Com isso, foram mantidos apenas os *tweets* em que ambos os anotadores concordavam quanto à polaridade, descartando *tweets* confusos ou irrelevantes. Essa prática reduziu o total de registros para **2831**, uma redução de quase 30%. Em seguida, os anotadores repassaram ao professor especialista os *tweets* anotados com menos confiança, para que fossem revisados.

Para se visualizar a fase de anotação, usou-se de técnica de representação para a criação da Figura 2. A cor rosa, representa os *tweets* de classe neutra. Na cor azul, os negativos e em amarelo, os positivos. Percebe-se, também, uma quantidade de *tweets* diferentes em cada classe, indicando desbalanceamento.

Figura 2 – Distribuição da Polaridade dos *tweets*



Fonte: Autoria própria (2019).

Mesmo com a anotação, ainda é necessário submeter os *tweets* à fase de **pré-processamento**. Essa fase é composta por 5 etapas: busca e substituição de padrões; tokenização; remoção de *stopwords*; *stemming* e; dicionário léxico. Essas etapas culminam na transformação dos tweets para um formato amigável ao classificador SVM.

A primeira etapa consiste em pesquisar por **padrões** nos *tweets*, partes do texto como datas, números, nomes de usuários, URLs e *hashtags*. Em seguida, esses padrões são substituídos por símbolos que representam cada um desses padrões.

Com isto, passa-se à fase de **tokenização**. Essa etapa, então, se encarrega de transformar um *tweet* em uma lista de palavras (*tokens*) que o compõem, removendo também a pontuação do *tweet*.

A etapa seguinte, por sua vez, seleciona as palavras mais significativas para a análise, removendo as menos importantes, ditas ***stopwords***. Na sequência, as palavras são substituídas por seus radicais, na etapa de ***stemming***, reduzindo em uma palavra as variações de plural, gênero etc.

Na última etapa, as palavras são aplicadas a um **dicionário léxico**. Esse dicionário é capaz de converter sentenças complexas em mais simples. Por exemplo, substituindo a sentença “reverte lucro” por “prejuízo”. Assim, quer-se

facilitar a análise de *tweets* que contenham expressões ou sentenças que conotem humor diferente do esperado.

Para a fase de análise, a base de dados de *tweets* foi separada em dois conjuntos. O primeiro, com 80% da base (2264 *tweets*), dedicado ao treinamento do SVM, usados para a aprendizagem do classificador. Já os 20% (566 *tweets*) restantes, dedicados à teste, usados para avaliar a assertividade do classificador.

No intuito de aumentar a robustez do classificador SVM, quatro abordagens (*kernels*) distintas foram testadas: linear; RBF; sigmoide e; polinomial. Cada *kernel*, então, foi submetido ao ***fine tuning***, técnica capaz de encontrar configurações ótimas, considerando o desbalanceamentos das classes.

No intuito de validar a contribuição das técnicas de busca e substituição de padrões e, também, dicionário léxico, foram realizadas três avaliações. A primeira, não aplicando as duas técnicas durante a fase de pré-processamento. Já a segunda, aplicando a busca e substituição de padrões, mas sem a aplicação do dicionário léxico. Por fim, a terceira, com a aplicação de ambas as técnicas.

Por fim, com a fase de visualização, é possível tornar essas avaliações mais claras por meio de tabelas. Assim como a geração de representações gráficas que auxiliam no entendimento da análise, vide Figura 2.

RESULTADOS

Na Tabela 1 resume-se a avaliação das três configurações descritas na seção anterior. Além disso, a Tabela 1 mostra os resultados referentes ao *kernel* RBF, avaliado como a melhor para as três abordagens.

A acurácia é utilizada como medida de avaliação do acerto do classificador. Já a precisão avalia, dos *tweets* preditos como corretos, quais definitivamente estavam corretos. O *recall*, entretanto, avalia quantos *tweets* deixaram de ser classificados corretamente. Em seguida, o F1 *Score*, uma média entre *recall* e precisão, ponderado para o pior resultado dentre os dois. Por fim, foi mensurado o tempo para classificar todos os 566 *tweets* do conjunto de teste, obtendo a média e desvio padrão, por meio de 30 iterações.

Tabela 1 – Avaliação das Três Abordagens

Abordagem	Acurácia (%)	Precisão (%)	*Recall* (%)	F1 *Score* (%)	Tempo (segundos)
Primeira	82,3322	82,5617	82,3322	82,2232	2,9267 ± 2,9191
Segunda	82,3322	82,7027	82,3322	82,3014	1,0925 ± 1,0874
Terceira	83,2155	83,4464	83,2155	83,1934	1,1298 ± 1,1253

Fonte: Autoria própria (2019).

Percebe-se na terceira abordagem, que faz uso das técnicas de busca e substituição de padrões e, também, do dicionário léxico, uma melhor pontuação nas métricas utilizadas. Além disso, como essas técnicas diminuem a complexidade da análise de dados, o processo de classificação se torna 2,6 vezes mais rápido em relação a primeira abordagem, a qual não as utiliza.

Dessa forma, com uma acurácia 83,2155%, alcança-se um resultado entre a assertividade humana, de 72% a 85% (NASCIMENTO et al., 2012), na classificação da subjetividade em textos.

CONCLUSÃO

Apesar do classificador SVM não superar a assertividade humana por 1,7845%, este é capaz de analisar milhares de *tweets* em segundos, ao passo que para o ser humano, levar-se-iam horas, ou até dias.

Tal resultado, entretanto, foi alcançado, sobretudo, pela execução de duas imprescindíveis fases: anotação e pré-processamento. Fases com a maior parcela de tempo investido neste trabalho de pesquisa.

Conjuntamente, a aplicação das duas técnicas propostas, não só trouxeram maior assertividade, bem como melhor performance para o classificador SVM. Possíveis, também, graças a criação da base de dados de *tweets*.

REFERÊNCIAS

BOLLEN, Johan; MAO, Huina; ZENG, Xiaojun. Twitter mood predicts the stock market. **Journal of computational science**, v. 2, n. 1, p. 1-8, 2011.

KANSAON, Daniel P.; BRANDÃO, Michele A.; DE PAULA PINTO, Saulo A. Análise de Sentimentos em *Tweets* em Português Brasileiro. In: **7° Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)**. SBC, 2018.

MARQUESONE, Rosangela. **Big Data**: Técnicas e tecnologias para extração de valor dos dados. São Paulo: Casa do Código, 2016. 202 p.

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.

STATISTA. **Leading countries based on number of Twitter users as of July 2019 (in millions)**. 2019. Disponível em: <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>. Acesso em: 05 ago. 2019.

STATISTA. **Only 34% of All Tweets Are in English**. 2013. Disponível em:<https://www.statista.com/chart/1726/languages-used-on-twitter/>. Acesso em: 11 ago. 2019.

ZAINUDDIN, Nurulhuda; SELAMAT, Ali. Sentiment Analysi Using Support Vector Machine. **IEEE 2014 International Conference On Computer, Communication, And Control Technology**. Langkawi, Kedah, Malaysia, p. 128-133. set. 2014.