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

BBLEX: um dicionário léxico multilíngue para o mercado de ações

BBLEX: a multilingual lexicon for stock market

Vítor Ângelo Misciato Teixeira
vitor.2018@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

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

RESUMO

Recentemente, com a ampla utilização das redes sociais, notícias e opiniões de usuários publicadas sobre o mercado de ações se tornaram conteúdo de grande valor preditivo para determinar o comportamento das ações. Com a velocidade e o volume que essas notícias e opiniões são geradas, a análise manual humana se torna inviável, sendo necessário o uso de ferramentas computacionais para o processamento dos textos, como dicionários léxicos. Entretanto, embora existam muitos dicionários de propósito geral, poucos são de domínio específico, com termos personalizados para o contexto do mercado de ações. Além disso, os que existem, em sua maioria, são limitados a um único idioma. Dito isto, este trabalho apresenta o BBLEX, um dicionário léxico multilíngue de domínio específico, desenvolvido com o objetivo de aumentar a assertividade da tarefa de Análise de Sentimento do grande volume de notícias publicadas, relacionadas ao mercado de ações brasileiro. As técnicas utilizadas para sua criação se mostraram promissoras, fazendo uso combinado das abordagens manual e automática. Nos testes realizados, o BBLEX mostrou-se mais eficiente e assertivo que os demais léxicos disponíveis utilizados para comparação, atingindo uma acurácia de 84,05% no melhor caso.

PALAVRAS-CHAVE: Redes sociais. *Big Data*. Processamento de linguagem natural.

ABSTRACT

RECENTLY, WITH THE WIDESPREAD USE OF SOCIAL MEDIA, NEWS AND USER OPINIONS PUBLISHED ABOUT THE STOCK MARKET HAVE BECOME CONTENT OF GREAT PREDICTIVE VALUE FOR DETERMINING STOCK BEHAVIOR. WITH THE SPEED AND VOLUME THAT THESE NEWS AND OPINIONS ARE GENERATED, HUMAN MANUAL ANALYSIS BECOMES IMPRACTICABLE, REQUIRING THE USE OF COMPUTATIONAL TOOLS FOR THE PROCESSING OF TEXTS, SUCH AS SENTIMENT LEXICONS. HOWEVER, ALTHOUGH THERE ARE MANY GENERAL-PURPOSE LEXICONS, FEW ARE DOMAIN SPECIFIC, WITH TERMS CUSTOMIZED FOR THE STOCK MARKET CONTEXT. IN ADDITION, THOSE THAT EXIST ARE MOSTLY LIMITED TO A SINGLE LANGUAGE. HAVING SAID THAT, THIS PAPER PRESENTS BBLEX, A DOMAIN SPECIFIC MULTILINGUAL LEXICON, DEVELOPED WITH THE AIM OF INCREASING THE ASSERTIVENESS OF THE SENTIMENT ANALYSIS TASK OF THE LARGE VOLUME OF PUBLISHED NEWS RELATED TO THE BRAZILIAN STOCK MARKET. THE TECHNIQUES USED FOR ITS CREATION PROVED TO BE PROMISING, MAKING USE OF COMBINED MANUAL AND AUTOMATIC APPROACHES. IN THE TESTS PERFORMED, BBLEX PROVED TO BE MORE

EFFICIENT AND ASSERTIVE THAN THE OTHER AVAILABLE LEXICONS USED FOR COMPARISON, REACHING AN ACCURACY OF 84.05% IN THE BEST CASE.

KEYWORDS: Social media. Big Data. Natural language processing.

INTRODUÇÃO

Nos dias atuais, com o avanço da *Web 2.0*, o volume de dados cresce cada vez mais. Marquesone (2016), à época, afirmava que cerca de 90% dos dados haviam sido gerados nos dois anos anteriores, em virtude da expansão da *Internet* e da ampla utilização dos dispositivos móveis, pelos quais grande parte dos usuários tem acesso aos meios de comunicação existentes. Redes sociais e *sites* como *Facebook*, *Instagram* e *Twitter* se tornaram ferramentas comuns do cotidiano das pessoas, onde expressam suas opiniões a respeito de produtos, serviços, organizações, entre outros (ALVES, 2015). A *International Data Corporation* (IDC) prevê que o volume de dados gerados cresça de 33 *Zettabytes* em 2018 para 175 *Zettabytes* até 2025, atingindo cerca de 59 *Zettabytes* até o final de 2020 (REINSEL; GANTZ; RYDNING, 2018). Grande parte desses dados, carregados de concepções dos usuários, são essenciais na tarefa de Análise de Sentimento e, assim, gerando informação e conhecimento.

A Análise de Sentimento, ou Mineração de Opinião, é o campo de estudo ligado ao Processamento de Linguagem Natural (PLN), que analisa os sentimentos, emoções e opiniões que os indivíduos têm sobre determinado assunto (LIU, 2012), geralmente classificados em um valor discreto, variando entre positivo, negativo ou neutro. De acordo com Taboada et al. (2011), as duas principais abordagens empregadas na extração de sentimentos são a abordagem lexical e o aprendizado de máquina. A primeira consiste no cálculo da orientação semântica dos textos com base nas palavras e sentenças presentes neles, caracterizando-se como um tipo de abordagem não supervisionada (TURNEY, 2002). Já o aprendizado de máquina supervisionado, envolve a criação de modelos de classificação a partir de textos rotulados (PANG; LEE; VAITHYANATHAN, 2002).

No contexto da análise de opiniões, um setor que sofre grande influência é o do mercado de ações. A capacidade dos usuários compartilharem notícias e classificações sobre ações e produtos uns com os outros, torna o seu impacto na bolsa de valores cada vez mais proeminente (LI et al., 2018). Com meios de comunicação tão rápidos e amplos, a informação, valiosa, chega com grande velocidade aos investidores. Contudo, para tomada de decisão, é necessária a análise de um grande volume de textos, o que demanda grande esforço para seres humanos (IM et al., 2013).

Dado o exposto, o objetivo deste trabalho é a criação de um dicionário léxico de domínio específico, visando aumentar a assertividade da tarefa de Análise de Sentimento do grande volume de notícias multilíngues publicadas relacionadas ao mercado de ações brasileiro.

MATERIAL E MÉTODOS

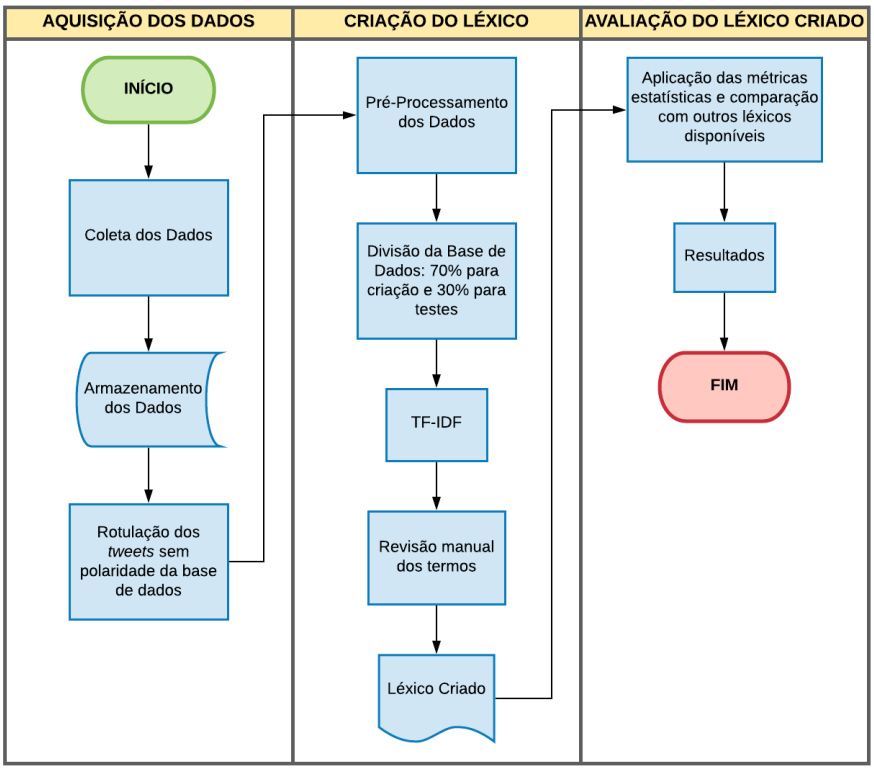
Para o desenvolvimento deste trabalho de pesquisa, foram utilizadas as seguintes tecnologias e ferramentas: Interface de Programação de Aplicações, *Application Programming Interface* (API) do *Twitter*; *Python*, como linguagem de programação; *MongoDB*, como Sistema de Gerenciamento de Banco de Dados (SGBD); Biblioteca *Natural Language Toolkit* (NLTK), para tarefa de PLN; Biblioteca *scikit-learn*, para aplicação das métricas estatísticas; *Spyder*, como Ambiente de

Desenvolvimento Integrado, *Integrated Development Environment* (IDE), voltado especialmente para programação científica em *Python*.

Para a criação do léxico, as atividades foram divididas em três etapas, sendo elas a Aquisição dos Dados, Criação do Léxico e Avaliação do Léxico Criado, que serão apresentadas a seguir.

Na Figura 1, é ilustrado um resumo das tarefas que compõem as etapas de criação do léxico.

Figura 1 - Fluxograma das etapas de criação do léxico



Fonte: Autoria Própria (2020).

Na etapa de Aquisição dos Dados, como método de criação do léxico, foi escolhido o método baseado em *corpus*, visto que se adequa ao padrão de resultado almejado, já que se faz o uso de um *corpus* de domínio específico, neste caso, sobre o mercado de ações. Assim, o *corpus* utilizado é composto por três *datasets* de *tweets*, cujos textos são oriundos do *microblog Twitter*, contendo até 280 caracteres.

O primeiro *dataset,* em português, foi criado durante o desenvolvimento desta iniciação científica, visando aumentar a quantidade de termos presentes no léxico. Para tanto, e com a ajuda de um *crawler* criado em *Python* e utilizando a API de busca oficial do *Twitter,* foram coletados *tweets* de perfis de veículos de informação especializados no mercado de ações, compreendidos entre 29 de Novembro de 2019 e 30 de Abril de 2020, armazenados no *MongoDB*. Como os *tweets* não possuem polaridade vinculada ao seu conteúdo, foi necessário um trabalho de rotulação manual. Com a ajuda de uma planilha eletrônica, foram rotulados em neutros, positivos e negativos, totalizando 2306 registros.

O segundo *dataset*, também em português, criado por Melo e Meinerz (2019), conta com 2831 entradas, e segue um padrão parecido com o descrito anteriormente para o primeiro *dataset*, com *tweets* compreendidos entre 16 de Novembro de 2015 e 16 de Maio de 2019, rotulados em neutros, positivos e negativos, apenas de notícias que mencionassem empresas listadas na bolsa de valores brasileira.

O terceiro *dataset*, em inglês, com *tweets* compreendidos entre 10 de Novembro de 2017 e 18 de Janeiro de 2020, foi retirado do *Kaggle*, uma comunidade *on-line* de Ciência de Dados, contendo 9470 manchetes de notícias das empresas listadas no índice norte-americano *Dow Jones*, com 58% (5482) dos *tweets* sendo positivos e 42% (3988) negativos.

Para a etapa de Criação do Léxico, o *corpus* foi dividido em dois conjuntos, levando em conta o idioma dos textos, português e inglês. Contudo, foram utilizados apenas os *tweets* rotulados como positivos ou negativos. Os *tweets* neutros foram desconsiderados, levando em conta que palavras que impactam diretamente na orientação semântica dos textos possuem uma polaridade, *a priori*, não neutra.

Como os *tweets* não possuem uma estrutura textual formal e contêm uma série de ruídos, foi necessária a aplicação de técnicas de pré-processamento. Dentre as etapas executadas nessa fase, estão:

a) Busca de substituição de padrões: remoção de elementos frequentes na maioria dos *tweets* e que não agregam valor na abordagem, como *links*, *emoticons*, *hashtags*, *cashtags* e números;

b) Conversão para minúsculo: facilita o tratamento das *strings*, padronizando-as para caixa baixa;

c) Tokenização: transformação do texto do *tweet* em uma lista de palavras (*tokens*) que o compõem;

d) Remoção de *stopwords* e pontuação: *stopwords* são palavras com alta frequência nos textos, mas que possuem pouco valor lexical e não contribuem para a análise, assim como a pontuação.

Após o pré-processamento dos dados, foi aplicado um esquema de validação de divisão nos dois conjuntos, onde 70% dos *tweets* serviram como conjunto de treinamento, para criação do léxico, e os outros 30% como conjunto de testes, para avaliação da técnica.

Como técnica para obtenção do léxico, foi escolhida a medida denominada *Term Frequency-Inverse Document Frequency* (TF-IDF), uma métrica estatística frequentemente utilizada em recuperação de informação e mineração de texto.

Após a lista com os termos e suas polaridades ser criada, foi aplicado um método de revisão manual, buscando a remoção de palavras contendo erros ortográficos ou que não são tão relevantes no contexto do mercado de ações. Como complemento do léxico, foram adicionados alguns termos retirados do glossário do *TradersClub*, uma plataforma para investidores do mercado financeiro.

A estrutura escolhida para o léxico foi o formato *Java Script Object Notation* (JSON), tendo como resultado uma lista de documentos, cada um contendo os

campos "termo" e "polaridade". Logo, o *MongoDB* foi escolhido como sistema de armazenamento, assim como ocorreu quando da coleta dos *tweets*. A Figura 2 demonstra como o léxico está estruturado.

Figura 2 - Estrutura do léxico criado

```
{'termo':'abaixo', 'polaridade':-1}
{'termo':'abaixo consenso', 'polaridade':-1}
{'termo':'abalando', 'polaridade':-1}
{'termo':'abandona', 'polaridade':1}
{'termo':'abilio', 'polaridade':1}
{'termo':'above', 'polaridade':1}
```

Fonte: Autoria Própria (2020).

Na etapa de Avaliação do Léxico Criado, com o objetivo de avaliar a eficácia do léxico criado, outros léxicos disponíveis também foram utilizados no conjunto de testes para comparação em cada idioma. Para o português, foram escolhidos o *SentiLex-PT* (SENT) (SILVA; CARVALHO; SARMENTO, 2012), como léxico de propósito geral, e o *Finance* (FIN) (PERES; VIEIRA; BORDINI, 2019), como léxico de domínio específico, enquanto para o inglês, optou-se pelo *Opinion Lexicon* (OL) (HU; LIU, 2004), como léxico de propósito geral, e pelo *Stock Market Lexicon* (SML) (OLIVEIRA; CORTEZ; AREAL, 2016), como léxico de domínio específico. Além dos léxicos disponíveis, também foi utilizado para comparação nos dois idiomas o Léxico Criado Sem Revisão (LCSR), o léxico criado antes do método de revisão manual ser aplicado, com o objetivo de destacar o impacto desse método no desempenho do léxico.

Como método para obtenção da polaridade do *tweet*, foi aplicado um somatório simples da polaridade das palavras de sentimento contidas no texto e tratamento de negação, invertendo a polaridade do termo caso outro de negação o precedesse. Sendo assim, classificou-se o texto em positivo, caso o valor obtido fosse maior ou igual a zero, e negativo, caso fosse menor que zero.

Como métricas de avaliação, foram aplicadas:

a) Acurácia: mede o total de textos classificados corretamente;

b) Precisão: dos textos que receberam um rótulo específico, indica quantos deles realmente estão corretos;

c) *Recall*: dos textos que compõem um rótulo específico, mostra quantos foram classificados corretamente;

d) F1-*Score*: uma média harmônica de precisão e *recall* para obter um único resultado, ponderando para o pior valor.

Para as medidas de precisão, *recall* e F1-*Score*, obteve-se o valor médio entre os obtidos para *tweets* positivos e negativos.

RESULTADOS E DISCUSSÃO

A seguir, são apresentados os resultados obtidos nos testes para a avaliação do léxico criado e desenvolvido, o *Bull-Bearish Lexicon* (BBLEX), um dicionário léxico multilíngue específico para o mercado de ações, composto por 2833 termos em português e inglês com suas respectivas polaridades anotadas.

A Tabela 1 contém os valores das métricas de avaliação, obtidos nos testes realizados no conjunto de textos em português.

Tabela 1 - Avaliação dos léxicos para o português

Léxico	Acurácia	Precisão	*Recall*	F1-*Score*
BBLEX	0.7955	0.7805	0.7445	0.7566
LCSR	0.7747	0.7480	0.7393	0.7432
SENT	0.6953	0.6509	0.6115	0.6153
FIN	0.5494	0.5159	0.5171	0.5145

Fonte: Autoria Própria (2020).

Percebe-se que os melhores valores foram obtidos pelo BBLEX, com uma acurácia de 79,55%, percentual 2,08% maior do que o obtido pelo LCSR, destacando o impacto da técnica de revisão manual no desempenho do léxico. Além disso, quando comparado aos léxicos SENT e FIN, o BBLEX mostrou-se mais eficiente, principalmente em relação ao FIN, com uma assertividade 24,61% maior.

Na Tabela 2, são apresentados os valores obtidos na abordagem para os textos em inglês, seguindo o mesmo padrão descrito na Tabela 1.

Tabela 2 - Avaliação dos léxicos para o inglês

Léxico	Acurácia	Precisão	*Recall*	F1-*Score*
BBLEX	0.8405	0.8411	0.8285	0.8331
LCSR	0.8109	0.8071	0.8012	0.8037
OL	0.6075	0.5955	0.5502	0.5166
SML	0.5417	0.5570	0.5571	0.5417

Fonte: Autoria Própria (2020).

Assim como para o português, os valores obtidos são melhores com o BBLEX, superando os demais léxicos em todas as medidas e atingindo uma acurácia de 84,05%. Como ocorreu para os testes em português, a técnica de revisão manual também aumentou a assertividade do BBLEX, com um valor de acurácia 2,96% maior em relação ao LCSR. Ademais, o OL e o SML tiveram um desempenho muito abaixo do BBLEX, com acurácias 23,3% e 29,88% menores, respectivamente.

CONCLUSÃO

Este trabalho de iniciação científica teve como objetivo a criação um dicionário léxico multilíngue de domínio específico, visando aumentar a assertividade da tarefa de Análise de Sentimento do grande volume de notícias multilíngues publicadas relacionadas ao mercado de ações brasileiro.

Os resultados obtidos foram analisados e avaliados quanto a eficácia do léxico criado. Constatou-se que o léxico BBLEX mostrou-se mais assertivo que os demais utilizados para comparação, atingindo uma acurácia de 84,05% no melhor caso.

Durante a pesquisa e descrição dos principais trabalhos correlatos, percebeu-se que a ausência de revisão manual em abordagens automáticas pode representar uma limitação no cálculo da orientação semântica dos textos, levando em conta que muitos termos podem não ter tanta relevância e, em contrapartida, o uso apenas da abordagem manual demanda muito tempo e esforço.

Dessa forma, destacam-se os seguintes aspectos que caracterizam as principais contribuições deste trabalho de iniciação científica:

a) Criação de um dicionário léxico multilíngue específico para o mercado de ações, composto por termos em português e inglês;

b) Uso combinado das abordagens manual e automática;

c) Expansão do *corpus* rotulado em português, para trabalhos futuros.

Entre os principais benefícios propiciados pelo BBLEX, destacam-se:

a) O aumento da taxa de assertividade na tarefa de Análise de Sentimento;

b) A possibilidade de ser utilizado em mais de um idioma.

Algumas limitações e dificuldades foram encontradas durante a realização do trabalho. As principais são listadas a seguir:

a) Identificação de fontes com dados previamente rotulados, o que requereu a rotulação manual de parte da base;

b) Escolha da melhor técnica para obtenção das polaridades dos termos que compõem o léxico;

c) Definição da estrutura final do léxico (quais informações conteria e em qual formato).

Após a realização da análise e avaliação dos resultados obtidos, constatou-se que este trabalho de pesquisa conseguiu alcançar o objetivo proposto de aumentar a assertividade na tarefa de Análise de Sentimento de notícias multilíngues relacionadas ao mercado de ações.

Para a continuidade deste trabalho, destacam-se abaixo as seguintes recomendações que se caracterizam como uma extensão natural ao BBLEX:

a) Expansão de termos presentes no léxico;

b) Adição de novas informações, como etiquetas morfossintáticas e formas flexionadas;

c) Desenvolvimento de um método para o cálculo da orientação semântica, específico para o mercado de ações.

A seguir, são apresentadas as referências bibliográficas utilizadas na elaboração deste trabalho.

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**. 2015. 133 f. Tese (Doutorado em Engenharia de Sistemas Eletrônicos e Automação) – Universidade de Brasília, Brasília, 2015. Disponível em: https://repositorio.unb.br/bitstream/10482/19345/1/2015_DeborahSilvaAlves.pdf. Acesso em: 27 mar. 2020.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04), 10, 2004, Nova Iorque. **Proceedings...** Nova Iorque: Association for Computing Machinery, 2004. p. 168-177.

IM, T. L. et al. Analysing Market sentiment in financial News sing lexical approach. In: 2013 IEEE Conference on Open Systems (ICOS), 2013, Kushing, Malaysia. **Proceedings…** Institute of Electrical and Electronics Engineers (IEEE), 2013. p. 145-149.

LI, Q. et al. Web Media and Stock Markets: A Survey and Future Directions from a Big Data Perpective. **IEEE Transactions on Knowledge and Data Engineering**, v. 30, n. 2, p. 381-399, fev. 2018.

LIU, B. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers, 2012.

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

MELO, J. G. S. de; MEINERZ, G. V. Análise de Sentimentos de Textos Voltados ao Mercado de Ações. In: Seminário de Iniciação Científica e Tecnológica da UTFPR (SICITE), 24, 2019, Pato Branco. **Anais...** Universidade Tecnológica Federal do Paraná, 2019.

OLIVEIRA, N.; CORTEZ, P.; AREAL, N. **Stock market sentiment lexicon acquisition using microblogging data and statistical measures**. Decision Support Systems, Elsevier BV, v. 85, p. 62-73, maio 2016.

PANG, B; LEE, L; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ACL-02 conference on Empirical methods in natural language processing (EMNLP '02), 2002. **Proceedings...** Estados Unidos da América: Association for Computational Linguistics. p. 79-86.

PERES, V.; VIEIRA, R.; BORDINI, R. Análises de Sentimentos: abordagem lexical de classificação de opinião no contexto mercado financeiro brasileiro. **In: Workshop**

of Artificial Intelligence Applied to Finance (WAIAF 2019). São José dos Campos: ITA, 2019.

REINSEL, D.; GANTZ, J.; RYDNING, J. **Data Age 2025**: The Digitization of the World – Fom Edge to Core. IDC White Paper, nov. 2018. Disponível em: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf . Acesso em: 7 jul. 2020.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a Sentiment Lexicon for Social Judgement Mining. In: International Conference on Computational Processing of the Portuguese Language (PROPOR), 10, 2012, Coimbra. **Proceedings...** Springer, 2012. p. 218-228.

TABOADA, M. et al. **Lexicon-Based Methods for Sentiment Analysis**, Computational Linguistics. MIT Press – Journals, v. 37, n. 2, p. 267-307, jun. 2011.

TURNEY, P. D. Thumbs up or thumbs down?. In: Annual Metting on Association for Computational Linguistics (ACL '02), 40, 2002. **Proceedings...** Estados Unidos da América: Association for Computational Linguistics, 2002. p. 417-424.