Análises de Sentimentos: abordagem lexical de classificação de opinião no contexto mercado financeiro brasileiro

Vitor Peres
Escola Politécnica - PUCRS
Porto Alegre, Brasil
Email: vitor.peres@edu.pucrs.br

Renata Vieira
Escola Politécnica - PUCRS
Porto Alegre, Brasil
Email: renata.vieira@pucrs.br

Rafael Bordini
Escola Politécnica - PUCRS
Porto Alegre, Brasil
Email: rafael.bordini@pucrs.br

***Abstract*—Financial news and Specialist Opinion bring us the latest information about stock market. Some studies have shown how information can be useful if it is analyzed correctly. Extracting sentiments and opinions from specifics texts, can be a way of assisting in decision-making. In this paper, we present a sentiment analyser for financial texts using lexicon-based approach, in portuguese. Using polarity lexicon, we can identify the positive or negative polarity of each term in the corpus. And also, we build a lexicon with a specific corpus from TradingView website.**

***Index Terms*—Sentimental Analysis; Bag-Of-Words; Lexicon; Precision; TF-IDF; Polarity; OpLexicon; Sentilex**

I. Introdução

Com o aumento da informação, gerada a partir do engajamento de usuários, vêm se observando a necessidade de geração de valor desses dados. No mercado financeiro não ocorre de maneira diferente, já que muitos dos investimentos passam por análises de especialistas com o objetivo de chegar a um senso comum de melhor opção.

Neste sentido, investidores que estão no inicio de carreira procuram por opiniões que transmitam as melhores ideias de investimento. Tendo em vista, que o domínio financeiro contém características linguísticas e semânticas únicas, cuja a interpretação depende de uma série de modelos semânticos que refletem nas ferramentas e estratégias utilizadas pelos especialistas [1], a análise de sentimentos chega como uma ótima opção. Através da análise de sentimentos é possível coletar qual o sentido da informação, categorizando-a em positiva ou negativa. Grande parte das aplicações que envolvem análises de sentimentos, vêm sendo direcionadas na análise de opiniões de produtos, onde o engajamento de consumidores servem de métrica para validar possíveis riscos e problemas [2].

Neste trabalho será apresentado a abordagem de análise de sentimentos utilizando dicionários léxicos no contexto financeiro. Além de responder a pergunta: léxicos de domínio geral, conseguem performar em textos de domínio específico (financeiro)? O objetivo deste trabalho será avaliar diferentes léxicos e qual a sua performance para textos de domínio específico. E por fim, a criação de um léxico de domínio financeiro, a partir de um corpus de opiniões de investidores, disponível na ferramenta TradingView[1].

Na seção II será apresentado a revisão literária em relação a análise de sentimentos. Na seção III, será apresentada o contexto e descrição dos dados deste trabalho. Na seção IV será apresentado o modelo desenvolvido. Na seção V serão apresentadas as etapas da construção do dicionário léxico. Seção VI os experimentos e resultados desta análise e por fim na seção VII a conclusão do trabalhos e as etapas futuras.

II. Revisão de Literatura

Por definição de Liu [3], a análise de sentimentos dedica-se ao tratamento de opiniões ou sentimentos que foram expressados em textos, e que através de um léxico de sentimento, que corresponde à uma orientação semântica ou a polaridade de palavras, assinalam quais sentidos textos se definem.

Mizumoto [4] propõe um estudo para determinar a polaridade de sentimentos em textos de noticias sobre mercado financeiro, utilizando dicionários léxicos. Além da construção de um dicionário léxico, fazendo o uso de aprendizado semi-supervisado, anotando manualmente a polaridade de um determinado numero de termos/palavras dos textos destas notícias. Ao final do estudo são comparados as polaridades determinadas pelo método proposta com as polaridade determinadas por especialistas de mercado financeiro.

San [5] também utiliza a abordagem léxica para a análise de sentimentos em notícias do mercado financeiro, sendo o mesmo feito sobre dois conjuntos de experimentos, que utilizam, ou não, o processo de stemming.

Palanisamy [6] através de léxicos, procuram-se identificar qual tipo de sentimento de tweets que contenham variações de palavras, emoticons e hashtags. Em [7], o autor propõe o emprego de coesão léxica baseada em textos, na detecção de sentimentos e polaridade de textos.

III. Contexto e Descrição de Dados

Com o aumento de pessoas investindo no mercado de ações, é possível observar que muitas delas ainda não possuem conhecimento total da movimentação dos preços das ações nos gráficos. Com isso o mercado acaba pagando,

[1]https://br.tradingview.com/

pois muitas ações acabam perdendo volume[2] por falta de negociadores ou até mesmo pela falta conhecimento para manejar ações ditas desinteressantes.

No mundo inteiro, o mercado de renda variável vem atraindo centenas de pessoas. Tendo essa capacidade de atração vinculada a duas características: ambiente completamente democrático, em que qualquer pessoa que possua capital terá um tratamento igual aos demais investidores, e principalmente a possibilidade de obter grandes ganhos praticamente do nada [8].

A. Base de Dados

Foram coletadas para este estudo, um série de dados que contemplam 3 fontes diferentes. Dessas 3 fontes podemos citar: **Trading View**, ferramenta que possibilita operar no mercado financeiro com auxílio da comunidade de investidores, e melhorar o desempenho a partir da observação de outros usuários. A **InfoMoney** é o maior site especializado em investimentos pessoais e educação financeira do Brasil, que fornece aos leitores informações que valem dinheiro de uma forma simples e agradável de ler. E por fim a **Investing.com** que é um portal financeiro global de dados e uma marca composta por 30 edições em 23 idiomas. Cada edição abrange uma ampla variedade de veículos financeiros locais e globais, incluindo ações, título públicos, commodities, câmbio, criptomoedas, taxas de juros, futuros e fundos.

A coleta de dados para a análise contém ao total 2024 textos. A base do TradingView, consiste de uma base de 615 textos, dentre esses dados, 435 textos possuem classe e 180 que não possuem, estes que são definidos pelo usuário da plataforma em 3 categorias; *Viés de Alta*, *Viés de Baixa* e *Educação* (exemplos disponíveis na tabela I). InfoMoney e Investing.com, possuem 1309 e 100 textos, respectivamente.

TEXTO	RÓTULO
Lucro da empresa ligado ao preço do petróleo e dólar . Dólar nos patamares de R$ 4,10 força a dívida da empresa para cima.	Viés de Baixa
GOLL4 passou por recente período de correção, e gora retoma sua tendência de alta, confirmada pela mudança de status do indicador HILO e maior volume. Possibilidade de upside de 38% aproximadamente.	Viés de Alta
GGBR4 esta em um canal de alta e formou uma congestão em retângulo na metade do canal, depois de uma perna de alta. Os preços tendem a atingir o topo do canal	Educação

TABLE I: **Textos e Rótulos Trading View**

IV. Modelo para Análise de Sentimentos

A. Pré-processamento de Dados

Para uma melhor performance do algoritmo criado, houve a necessidade de realizar a limpeza dos dados, que possibilita através de abordagem de dicionario léxico percorrer as sentenças de maneira simples. Esse processo têm sido aplicado a textos que não possuem a estrutura formal, no caso textos de Twitter, e contem uma série significativa de ruídos [9]. Estes que pela definição, se caracterizam por dados que não entregam informação útil para análise em questão. Dentre as etapas de pré-processamento segue:

- *Remoção Caracteres Especiais*: caracteres especiais, que dentro deste contexto, como quebra de linhas, url's e cifrões não geram valor na análise e sua retirada é eminente;
- *Conversão para Minúsculos*: o processo de conversão do texto para minúsculo, melhora a combinação de palavras e reduz o problema de dimensionalidade [10];
- *Remoção de Números e Pontuações*: a presença de pontuação em alguns casos, denota a existência de algum tipo de sentimento. No caso de uma frase que tem seu termino com uma *exclamação*, pode induzir a um sentido de intensidade positiva ou negativa. O mesmo vale para números, que indicam quantidade de determinado termo e não indicam sentimento [11];
- *Remoção StopWords*: as StopWords são palavras que possuem alta frequência nas sentenças, e não contém valores úteis dentro da análise. São classificadas como StopWords: preposições, artigos, conjunções e dentre outros. Nesta análise, foi feito o uso do pacote NLTK de StopWords em português [12].

B. Polaridade de Textos

A classificação dos textos parte de uma comparação das sentenças (textos) com as palavras disponíveis no léxico. Após a fase de pré-processamento cada sentença é *Tokenizada*, processo que separa em objetos únicos os termos presentes na string, facilitando a análise por parte do algoritmo [13].

A partir desta etapa, são comparadas cada objeto único das strings com os termos disponíveis em cada léxico. A cada palavra encontrada, é adiciona a uma lista temporária os valores referentes a positivo (1) e negativo (-1), e caso não haja o termo a comparar no léxico é adicionado o valor neutro (0). Após, somas são realizadas para concretizar a polaridade dos textos, conforme a fórmula a seguir:

[2]Volume representa o nível de atividade do mercado, sendo o motor que empurra e sustenta os movimentos dos preços. Baixo volume demonstra baixa participação dos investidores, e portanto pequeno comprometimento financeiro com o movimento em si, tornando-se frágil. Alto volume demonstra que os investidores estão atentos e ativos, oferecendo sustentação e validade aos movimentos aos sinais que percebemos [8]

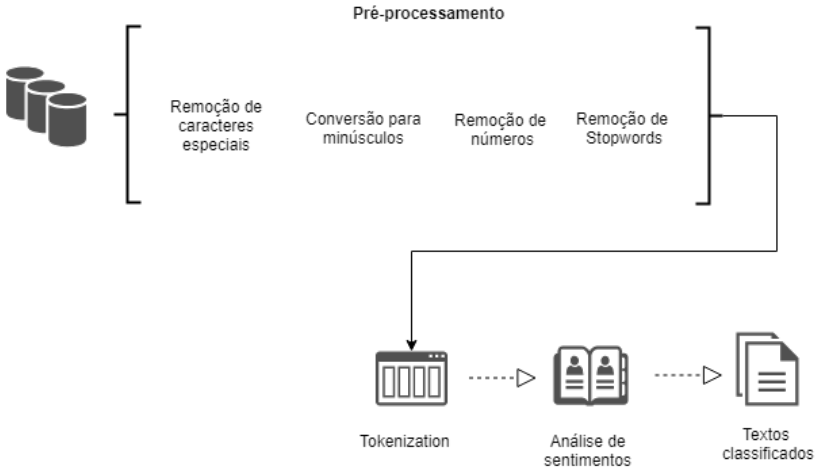

Fig. 1. Modelo proposto para análise e classificação de sentimentos.

$$\sum_{k=1}^{N} t(i,j)$$

Onde *t(i, j)* representa a palavra/termo (i) e sua polaridade (j) de acordo com a sua disponibilidade no léxico. Realizado somatório dos termos de uma sentença, são analisados em qual pesos devem ser classificados. Por exemplo, um texto tem como resultado do seu somatório o valor de 4, visto nossa caracterização anterior de termos, classificamos em *Muito Positivo*, conforme a tabela II.

Valor Somatório	Classe
menor que -2	Muito Negativo
entre -0.5 e -1	Negativo
entre -0.5 e 0.5	Neutro
entre 0.5 e 1	Positivo
maior que 2	Muito Positivo

TABLE II: Tipos de Classificação

C. Abordagem Lexical

1) OpLexicon: construído a partir do estudo de Souza [14], consiste de uma aplicação de 3 técnicas diferentes presentes na literatura: *Turney's Corpus-based*, *Thesaurus-based* e uma variação de *Mihalcea*, utilizando um sistema de tradução automática.

O corpus utilizado no estudo é composto por 346 reviews de filmes em português brasileiro, extraídos do site CinePlayers e Cinema com Rapadura, e de 970 textos jornalísticos sobre diferentes temas extraídos do PLN-Br Categ corpus. TEP Thesaurus foi utilizado como recurso léxico dentro do estudo, este que contém 44077 palavras e anotações de synsets e antônimos.

2) SentiLex: constitui de 6321 lemmas e 25406 formas flexionadas. Os atributos de cada entrada do léxico são: a polaridade do adjetivo, alvo do sentimento e método de atribuição de polaridade.

3) UniLex: foi construído a partir de um corpus extraído do twitter, sobre o tema política, rotulados e comparados para a criação do léxico [15].

V. Construção de Dicionário Léxico

Verificado a disponibilidade de dados rotulados (TradingView), observou-se a possibilidade da criação de um léxico especifico em mercado financeiro. Como mencionado anteriormente, o dataset extraído da ferramenta TradingView, contém a classificação do texto inserida manualmente pelo usuário. O processo será abordado utilizado duas técnicas comuns em processamento de linguagem natural: Bag of Words e TF-IDF.

A. Bag Of Words

A essência deste técnica é converter documentos de textos em vetores, de modo que cada documento seja convertido em um vetor que represente a frequência de todas as palavras distintas presentes em um espaço vetorial de um documento [16].

B. TF-IDF

Abreviação do termos *Term Frequency-inverse Document Frequency*, que indica a ponderação de termos que ocorrem em termos proporcionalmente à sua frequência [16]. Esta técnica foi originalmente desenvolvida como uma métrica para funções de classificação, na busca de resultados em mecanismos de busca, baseadas em consultas de usuários.

$$W_{t,d} = \left(\frac{Freq_{t,d}}{Max}\right) * log_2\left(\frac{N}{n_t}\right)$$

Onde $W_{t,d}$ é o peso de termo t no documento d. $Freq_{t,d}$ é a quantidade de vezes que o termo/palavra ocorre em um texto. Max é o termo que possui maior ocorrência no texto. N é o total de textos na base de testes, n_t é o número de textos na base de testes que possui o termo t.

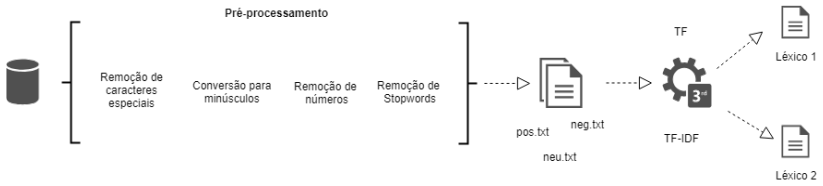

Fig. 2. Modelo Construção do Léxico.

VI. Experimentos e Resultados

A. Análise de Sentimentos

Neste trabalho iniciamos os experimentos utilizando o modelo anteriormente descrito, nos 3 conjuntos de dados: *TradingView, Investing.com e InfoMoney*. Inicialmente foi realizado os experimentos sem nenhum tipo de tratamento dos dados, e assim foi observado métricas fora do normal. Então novamente, verificou o estado da arte e verificado quais seriam as etapas fundamentais para um melhor resultado, com os mesmos presentes na tabela IV.

Com o objetivo de avaliar o modo que o algoritmo esta classificando os textos, foi definido a utilização das métricas Precision, Recall, F1-score e Acuracy, bastante utilizando em algoritmos de classificação supervisionada [17]. Neste etapa, utilizaremos o conjunto de dados do TradingView, que do montante total de 615 textos, apenas 435 são rotulados. Estes sendo 286 positivos, 19 neutros e 130 negativos.

Métrica	OpLexicon	SentiLex	UniLex
Precision	0,693	0,721	0,612
Recall	0,763	0,822	0,631
F1-score	0,725	0,76	0,621
Acuracy	0,6	0,673	0,502

TABLE III: Performance do Algoritmo Léxico

Dentre os léxicos que melhor obtiveram acurácia podemos notar que o SentiLex chegou a **0.673** contra **0.6** do OpLexicon. Também vale ressaltar que nas outras métricas de Recall, Precision e F1-score, SentiLex teve pequena vantagem em relação ao OpLexicon (tabela III).

B. Composição do Léxico

Feita esta análise, partimos agora para a construção de um léxico, este que foi baseado na técnica TF-IDF, definida na seção anterior. O processo de construção inicia no pré-processamento de dados, e na divisão do dataset em arquivos de acordo com sua classificação prévia (Figura 2). O uso de *Bag of Words* ajuda na conversão de documentos em vetores [16], e facilita a manipulação dos mesmos, além de possibilitar a geração da frequência do aparecimento de palavras. De cada arquivo separado, foram indicados a frequência assim como sua polaridade.

- **Positivo**: "*Alta, Queda, Oportunidade, Rompendo*";
- **Neutro**: "*Força, Retração, Romper, Risco, Queda*";
- **Negativo**: "*Vermelho, Resistência, Negativo*".

Fica evidente que apenas calcular a frequência de palavras pode não ser a melhor opção, sendo assim optamos por verificar o mesmo processo, porem com a aplicação da técnica TF-IDF, presente também na construção do léxico Unilex [15]. Isso também é um estimulo já que muitas das palavras se repetiam com polaridades diferentes.

- **Positivo**: "*Sólido, Favorável, Alto, Alcançar*";
- **Neutro**: "*Acentuado, Especialista, Permanecer*";
- **Negativo**: "*Desastre, Invertido, Impedir*".

Com o uso de TF-IDF verificou-se alguns termos que no significado real da palavra, estariam relacionados ao sentimento contrario. Por exemplo, foram classificados como positivos os termos **Reagiu** e **Caiu**, que no contexto financeiro, o termo "reagiu" pode estar relacionado a uma opção de investimento que não fosse trazer grande lucro, e o termo "caiu", à um acontecimento que pode gerar lucro devido a sua variação de preço.

C. Análise com léxico

Por fim, foi realizado o treinamento do nosso léxico, em cima dos dados do TradingView que não haviam rótulos, possibilitando assim verificar que tipo de resultado podemos obter utilizando um léxico de conteúdo especifico. Realizamos o procedimento mencionado na seção IV e compilamos os resultados dos léxicos SentiLex, OpLexicon e Lexico Financeiro na figura 3.

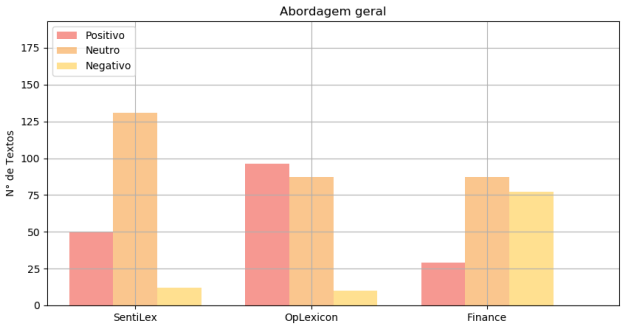

Fig. 3. Comparação léxicos.

Novamente, com o objetivo de estimar o quão ótimo nosso léxico classificou os textos, voltamos para o montante de 435 dados rotulados do TradingView. Nesse caso, os valores de **0.78**, **0.64**, **0.69** e **0.62** foram obtidos para os casos

	SentiLex			OpLexicon			UniLex		
	POS	**NEU**	**NEG**	**POS**	**NEU**	**NEG**	**POS**	**NEU**	**NEG**
InfoMoney	174	999	135	343	649	316	531	247	527
Investing.com	5	86	9	17	60	23	40	27	32
TradingView	220	157	71	390	112	124	388	49	195

TABLE IV: Resultados das análises de sentimentos por léxico.

de Precision, Recall, F1-score e Acurácia, respectivamente. Um fato a se observar é a relação de textos negativos classificados pelo léxico criado, já que trata-se de textos que não haviam sido classificados por parte dos usuários, pode se considerar que a opinião dos mesmos eram de pessimismo em relação a determinada ação, e que com um léxico específico podemos obter uma melhor leitura da análise.

D. Expansão do Léxico

Com o objetivo de enriquecer o léxico, foi adicionada as features de Stemming e POS Tag (Part of Speach). Esse processo foi realizada em duas partes: a primeira na identificamos o POS utilizando o léxico que melhor perfomou, no caso o SentiLex. E a segunda parte, utilizando a técnica Stemming, que consiste em reduzir a palavra sua raiz (ou radical), buscamos outras variações das palavra/termos disponível no léxico SentiLex. Sendo o léxico disponível no GitHub[3] do projeto.

VII. Conclusão

Este trabalho apresentou o uso de dicionários léxicos, com o objetivo de classificar textos específicos sobre o mercado financeiro. Além de verificar o quão próximo um léxico pode classificar textos já classificados.

Dos experimentos realizados, verificamos que o SentiLex conseguiu obter o melhor resultado comparado com o OpLexicon, tendo como resultado de precisão mais de 70%. Já o processo de construção de um léxico, foi analisado duas abordagens, sendo a ultima, utilizando TF-IDF, a que melhor se aproximou dos termos que o mercado de ações contém. E que se utilizados de maneira conjunta podem trazer resultados ainda mais expressivos.

Como trabalhos e extensões futuras desta pesquisa, é citada a possibilidade de comparação de algoritmos supervisionados como NaiveBayes, SVM e dentre outros, com a abordagem léxica. Outra possibilidade, é a inclusão de mais técnicas de pré-processamento de dados, como por exemplo o uso de Stemming e Normalização de textos, que podem ainda assim influenciar em melhores resultados. E na construção do léxico, a inserção de Part-of-Speach nos termos, isso sendo possível através da integração com outros léxicos mais completos, como a WordNet por exemplo. Além de inserir as variações gramaticais das palavras, utilizando a técnica stemming de maneira reversa.

[3]https://github.com/viitormiiguel/AnalysisFinancial

References

1 Maia, M., Freitas, A., and Handschuh, S., "Finsslx: A sentiment analysis model for the financial domain using text simplification," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Jan 2018, pp. 318–319.

2 Shahana, P. and Omman, B., "Evaluation of features on sentimental analysis," *Procedia Computer Science*, vol. 46, pp. 1585 – 1592, 2015, proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace Island Resort, Kochi, India. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915001520

3 Zhao, J., Liu, K., and Xu, L., "Sentiment analysis: Mining opinions, sentiments, and emotions bing liu (university of illinois at chicago) cambridge university press, 2015 isbn 9781107017894," *Computational Linguistics*, vol. 42, pp. 1–4, 06 2016.

4 Mizumoto, K., Yanagimoto, H., and Yoshioka, M., "Sentiment analysis of stock market news with semi-supervised learning," in *2012 IEEE/ACIS 11th International Conference on Computer and Information Science*, May 2012, pp. 325–328.

5 Im, T. L., San, P. W., On, C. K., Alfred, R., and Anthony, P., "Analysing market sentiment in financial news using lexical approach," in *2013 IEEE Conference on Open Systems (ICOS)*, Dec 2013, pp. 145–149.

6 palanisamy, P., Yadav, V., and Elchuri, H., "Serendio: Simple and practical lexicon based approach to sentiment analysis," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, 2013, pp. 543–548. [Online]. Available: http://aclweb.org/anthology/S13-2091

7 Devitt, A. and Ahmad, K., "Sentiment polarity identification in financial news: A cohesion-based approach," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, 2007, pp. 984–991. [Online]. Available: http://aclweb.org/anthology/P07-1124

8 Bisi, T., "Análise técnica," in *L&S Educação - Traders ensinando traders*, Educação, L., Ed., 2017.

9 Symeonidis, S., Effrosynidis, D., and Arampatzis, A., "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," *Expert Systems with Applications*, vol. 110, pp. 298 – 310, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417418303683

10 dos Santos, C. and Gatti, M., "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, 2014, pp. 69–78. [Online]. Available: http://aclweb.org/anthology/C14-1008

11 He, Y., Lin, C., and Alani, H., "Automatically extracting polarity-bearing topics for cross-domain sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 123–131. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002489

12 Loper, E. and Bird, S., "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 63–70. [Online]. Available: https://doi.org/10.3115/1118108.1118117

13 II, T. B., *Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing*. Apress, 2018.
14 Souza, M., Vieira, R., Busetti, D., Chishman, R., Alves, I. M., and Unisinos, F. D. L., "Construction of a portuguese opinion lexicon from multiple resources," in *In 8th Brazilian Symposium in Information and Human Language Technology - STIL, Mato Grosso*, 2011.
15 Karine França de Souza, D. H. D., "Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro." PUC Minas, 2017.
16 Sarkar, D., *Text Analysis with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data*. Apress, 2016.
17 Olson, D. and Delen, D., *Advanced Data Mining Techniques*, 01 2008.