DESENVOLVIMENTO E COMPARAÇÃO DE LÉXICO E REDE NEURAL PARA A ANÁLISE DE SENTIMENTO DE NOTÍCIAS

Guilherme Martins Marques
Graduando em Análise e Desenvolvimento de Sistemas, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Cubatão, SP, Brasil.

Glauber R. Colnago
Professor Doutor, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Cubatão, SP, Brasil.

Juan A. Castañeda-Ayarza
Professor Doutor, Pontifícia Universidade Católica (PUC), Campinas, SP, Brasil.

Resumo: A análise de sentimentos visa a classificação da positividade/negatividade de textos, sendo um tópico da Mineração de Textos (*Text Mining)*, onde são empregadas técnicas para extrair informações importantes de textos, possuindo aplicações como em redes sociais, seções de comentários e avaliações de produtos ou serviços como auxílio à tomada de decisões de empresas. Este trabalho focou na análise de notícias, sendo usados dois métodos: léxico e rede neural. O projeto envolveu a elaboração e avaliação de um léxico de sentimento e o treinamento de uma rede neural recorrente com LSTM (*long short-term memory*) na categorização do sentimento de notícias. O léxico atuou melhor em situações onde a distinção entre positivo/negativo poderia ser identificada através de termos isolados, mas falhou em situações que necessitavam de mais contexto. Já a rede neural apresentou no geral melhores resultados que o léxico, conseguindo também ser mais consistente em suas previsões.

Palavras–chave: Mineração de Textos. Léxico de sentimento. Redes neurais artificiais.

Abstract: Sentiment analysis aims to classify the positivity/negativity of texts, being a topic of Text Mining, where techniques are used to extract important information from texts, having applications such as in networks social, commentary sections and evaluations of products or services as an aid to business decision-making. This paper focused on the analysis of news, using two methods: lexicon and neural network. The project involved the elaboration and evaluation of a lexicon of feeling and the training of a recurrent neural network with LSTM (long short-term memory) in the categorization of the news feeling. The lexicon acted best in situations where the distinction between

positive/negative could be identified through isolated terms, but failed in situations that needed more context. The neural network, however, presented better results than the lexicon, and was also more consistent in its predictions.

Keywords: Text Mining. Sentiment lexicon. Artificial neural network.

INTRODUÇÃO

Uma área muito utilizada em aplicações de Mineração de Textos é a de Processamento de Linguagem Natural (PLN, ou NLP, *Natural Language Processing*, em inglês). Esta área trata de problemas relacionados com a compressão e extração de informações a partir de documentos em linguagem natural, ou seja, textos escritos (FREITAS, 2014). Durante muito tempo este tipo de análise procurava somente a identificação de informações factuais (quem, o quê, onde, quando etc.), porém com o tempo o interesse se voltou também para análise das informações subjetivas que os textos convêm, trazendo a análise de sentimento ou mineração de opiniões (PANG;LEE, 2008).

Em relação ao NLP, parte fundamental para sua aplicação está no uso de léxicos. Neste contexto, léxicos são componentes de um sistema computacional que guardam informações relacionadas a palavras ou expressões, contendo dados semânticos e/ou gramaticais. Desta forma, surgiram também léxicos próprios para a análise de sentimento, que guardam informações sobre a polaridade de palavras ou expressões em seu teor positivo e/ou negativo (FREITAS, 2014).

Uma outra área muitas vezes utilizada em conjunto com a mineração de textos é a de aprendizagem de máquina (*machine learning*, em inglês). Esta área é um ramo da inteligência artificial que trata de sistemas capazes de adquirir conhecimento a partir de dados. Na área de mineração de textos, uma de suas aplicações é a classificação textual (ROLIM; FERREIRA; COSTA, 2017).

Os métodos de aprendizagem de máquina podem ser divididos em duas formas básicas, aprendizado supervisionado e aprendizado não-supervisionado (FERNEDA, 2006). O primeiro necessita conhecimento prévio do comportamento desejado enquanto o segundo não. Por exemplo, considerando uma classificação de textos, se

já é conhecida as categorias em que os textos devem ser divididos, técnicas de aprendizado de máquina podem ser usadas para o treinamento em cima de um conjunto de textos já classificados. O algoritmo então pode começar a prever a categoria de novos textos, usando os conhecimentos conseguidos no treinamento. Este tipo de aprendizado seria supervisionado. Por outro lado, se não existem exemplos de classificação, tratamos de um aprendizado não-supervisionado, ou seja, o algoritmo deverá agrupar automaticamente os textos que julgar semelhantes.

Com o objetivo de classificação, existem várias técnicas de aprendizagem de máquina, como: Regressão, Árvores de Decisão, Naive Bayes, Redes Neurais, entre outras.

Este trabalho tem como objetivo o uso de técnicas de mineração de textos na criação de um léxico de sentimento e no treinamento de uma rede neural, fazendo a comparação do desempenho das duas para posterior uso em análises de sentimento envolvendo notícias.

FUNDAMENTAÇÃO

Existem várias técnicas para a criação de léxicos, porém no geral elas podem ser agrupadas em três formas principais: manual, usando um dicionário e usando um conjunto de documentos (LABILLE, GAUCH e ALFARHOOD, 2017).

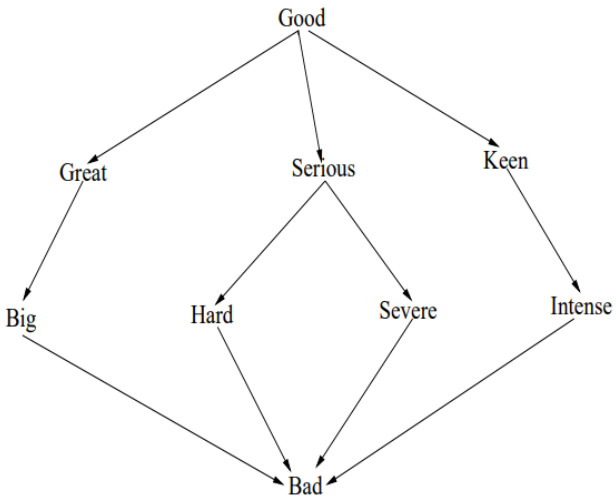
Manual: Esta forma se refere a polarização manual de uma grande quantidade de palavras e/ou expressões. Este tipo de classificação pode ser facilitado usando um grande grupo de pessoas. Serviços como o *Mechanical Turk* (https://www.mturk.com/), que permitem a contratação de mão-de-obra em demanda para trabalhos que precisam de inteligência humana, podem ser usados para facilitar o processo.

Um exemplo seria uma pesquisa realizada na Universidade de Vermont, onde pesquisadores fizeram uma análise de sentimento (que chamaram de análise de “alegria”) na rede social *Twitter* usando um léxico próprio. Durante 3 anos foram recolhidos cerca de 4,5 bilhões de *tweets*, contabilizando 46 bilhões de palavras. Dentre estas foram selecionadas 10.222, que foram votadas numa escala de 1 a 9

(sendo 1 menos alegre e 9 mais alegre) por usuários do *Mechanical Turk* (DODDS *et al.*, 2011). Dessa forma, conseguiram criar um léxico para uso em suas análises.

Usando um dicionário: Este método assume que já se possui um léxico confiável, porém reduzido. O léxico então é expandido usando sinônimos e antônimos. Essa expansão pode ser feita recursivamente consultando um dicionário de sinônimos, porém o problema deste método é que a coerência de sinônimos tende a cair com a distância. Usando o léxico de sinônimos do *WordNet*, um dicionário computacional da língua inglesa, já existem pelo menos 4 maneiras de chegar da palavra '*good*' para a '*bad*', antônimos, usando apenas três pulos de sinônimos (Figura 1).

Figura 1 – Quatro maneiras de chegar de good para bad em três pulos



Fonte: Godbole, Srinivasaiah e Skiena (2007)

Para amenizar estes problemas, Godbole, Srinivasaiah e Skiena (2007) apontam algumas medidas que tomaram na criação de seu algoritmo:

- Associar a polaridade para cada palavra encontrada, mesma polaridade se for um sinônimo, polaridade oposta se for um antônimo.
- Como a confiança diminui a cada pulo, a significância de uma palavra segue uma função que leva em conta sua profundidade (o número de pulos que levou

até chegar a essa palavra). Dessa forma, sua significância diminui exponencialmente quanto mais longe da palavra original. Para calcular o valor final da palavra, soma-se ou calcula-se a média dos valores obtidos em todos os caminhos.

- Ainda há outros métodos, como a implementação de uma segunda iteração que leva em conta quantas vezes os caminhos alternaram entre polaridade positiva e negativa, considerando somente aqueles mais confiáveis que não tiveram tantas alternações.

Dessa forma, o algoritmo utilizado em sua pesquisa conseguiu gerar mais de 18.000 palavras dentro dos primeiros cinco pulos a partir do pequeno conjunto de palavras que começaram (GODBOLE, SRINIVASAIAH e SKIENA, 2007).

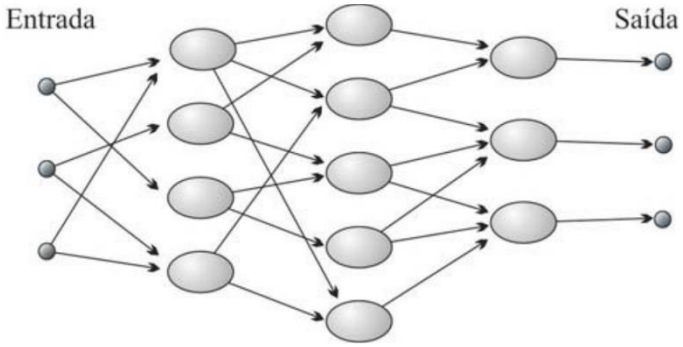
No caso da língua portuguesa, ainda existe uma outra forma de expandir léxicos. Comparado com a língua inglesa, o português possui mais variações de conjugação de verbos e tempo. Assumindo-se que o índice atribuído para uma forma do verbo se repetiria de forma parecida para as outras formas verbais, pode-se fazer a conjugação dos verbos dentro do léxico inicial usando os mesmos valores.

Usando um grupo de documentos: Outra forma é usando um grupo de documentos já valorados com índices de sentimento, as palavras então são valoradas segundo os índices dos documentos que mais aparecem. Este método é bem variado, podendo fazer desde uma média ponderada de índices até a aplicação de treinamento em *Machine Learning*. Um exemplo seria o léxico criado por Freitas (2014) a partir do *corpus ReLi*, um conjunto de resenhas de livros publicadas na internet, incluindo textos de caráter informal e formal. Do *corpus* foram coletadas 200 resenhas de cada um dos treze livros disponíveis. Foi feita automaticamente a classificação das palavras e manualmente a anotação quanto a expressão de opinião, onde foram marcadas a polaridade de cada frase e de alguns segmentos que expressavam opinião. No geral, a criação deste léxico foi mais manual por conta da preocupação dos autores de deixar somente palavras que expressavam opinião na maioria dos contextos, mas que exemplifica o uso de um *corpus* de documentos.

Neste trabalho, foi usada a técnica descrita como “usando um dicionário” para a criação de um léxico, e a conjugação de verbos foi a forma escolhida de expansão do mesmo.

Em relação às redes neurais, elas são modelos computacionais inspirados no funcionamento simplificado do cérebro humano. Uma rede neural artificial pode ser vista como um grafo (Figura 2), onde os nós representam os neurônios e as ligações entre eles exercem a função de sinapses (FERNEDA, 2006). O que é modificado durante o treinamento de uma rede neural são os pesos relativos a cada ligação.

Figura 2 – Representação simplificada de uma rede neural artificial



Fonte: Ferneda (2006)

Em áreas que envolvem o processamento de linguagem natural, são muito usadas as redes neurais recorrentes. São capazes de processar dados em sequência, assim como usar informações contextuais ao mapear sequências de entrada e saída. Sua operação em *loops* permite um *feedback* constante dos processamentos anteriores, armazenando informações ao processar novas entradas.

Porém, uma rede neural recorrente convencional pode não funcionar muito bem para o processamento de sequências muito longas, que exigem a persistência de informações de longo prazo. Para resolver esse problema, existe uma variação de sua arquitetura, a LSTM (*Long Short-Term Memory*). Esta variação se utiliza de um mecanismo específico nas camadas ocultas da rede, chamado de “células de memória”, um conjunto de subestruturas conectadas de maneira recorrente, blocos de memória que usam diferentes “válvulas” (de entrada, de saída e de esquecimento),

para o gerenciamento da memória. Para a realização deste trabalho foi usada uma rede neural recorrente com LSTM.

MATERIAIS E FERRAMENTAS

Para a manipulação e o armazenamento dos dados foi usado o banco *Microsoft SQL Server 2016 Express*, juntamente com o *Microsoft SQL Server 2016 Management Studio Express*, para o gerenciamento das instâncias do *SQL Server*. (MICROSOFT).

Foi usado o editor de planilhas *Microsoft Office Excel* na análise e criação de gráficos, além da criação de macros para automação de uma das etapas da expansão do léxico que envolveu a conjugação de verbos.

No treinamento e manipulação da rede neural foi utilizada a linguagem de programação Python (Python). A rede neural foi desenvolvida por um pesquisador em Ciência da Computação na Universidade Federal de Pernambuco, usando a linguagem Python em conjunto com pacotes como Keras e TensorFlow (bibliotecas de redes neurais e aprendizado de máquina). O código foi obtido a partir de seu GitHub (https://github.com/luisfredgs).

Para o tratamento e análise dos dados, usou-se a ferramenta open source R (THE R FOUNDATION, 2017), uma linguagem e ambiente de desenvolvimento integrado para computação estatística e gráfica, juntamente com o R Studio (RSTUDIO TEAM, 2016), nos provendo um editor para o código R, assim como *debugging* e ferramentas de visualização. Os principais pacotes desta ferramenta usados neste projeto seguem:

- *lexicon* – uma coleção de léxicos e listas de palavras para uso em *Text Mining*, incluindo a *SentiWordNet*, usada neste trabalho.
- *RYandexTranslate* – oferece uma interface entre o R e a API do '*Yandex Translate'* (https://translate.yandex.com/), um sistema *web* de tradução de textos.
- RODBC – oferece uma interface com a ODBC (*Open Database Connectivity*), permitindo o acesso direto ao banco de dados do *SQL Server* pelo R.

Como dicionário base para criação do léxico utilizou-se o *SentiWordNet* (http://sentiwordnet.isti.cnr.it/), um léxico criado para mineração de opinião em inglês. Essa ferramenta se baseia no dicionário léxico *WordNet* (https://wordnet.princeton.edu/), onde palavras são agrupadas em conjuntos de sentidos chamados *synsets*. Para cada *synset*, o *SentiWordNet* associa três valores de pontuação, positivo, negativo e objetivo, que foram obtidos utilizando um método de aprendizagem de máquina semi-supervisionado.

A Tabela 1 mostra a configuração dos dados neste léxico: POS (“part of speech”, divisão em aditivos, substantivos, verbos, etc.) e ID como identificadores de *synsets*, *PosScore* e *NegScore* representando os valores de positividade e negatividade, *SynsetTerms* listando os termos e seus números de sentido, e *Gloss* descrevendo o sentido das palavras seguido de exemplos de seu uso.

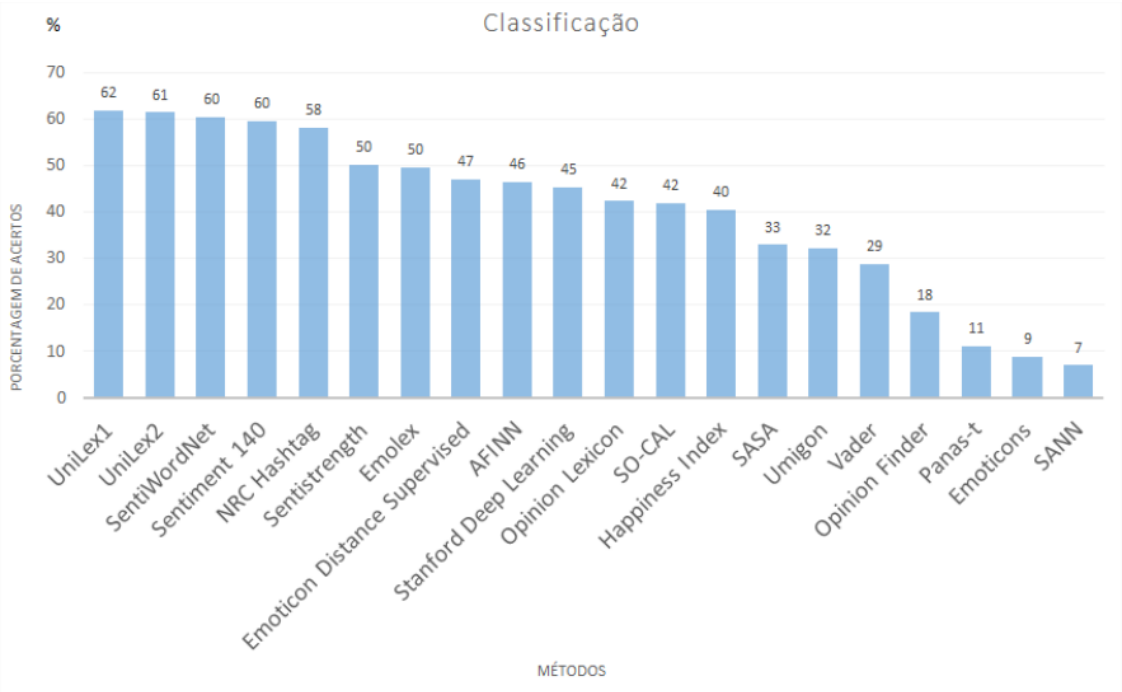
Tabela 1 - Exemplo da organização de dados no SentiWordNet

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	1740	0.125	0	able#1	(usually followed by `to') having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at last able to buy a car"; "able to get a grant for the project"
a	2098	0	0.75	unable#1	(usually followed by `to') not having the necessary means or skill or know-how; "unable to get to town without a car"; "unable to obtain funds"
a	2312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ or organism; "the abaxial surface of a leaf is the underside or side facing away from the stem"
n	662340	0.125	0.25	chemotherapy#1	the use of chemical agents to treat or control disease (or mental illness)
n	662527	0.25	0.125	correction#7	treatment of a specific defect; "the correction of his vision with eye glasses"
n	662972	0.25	0.25	insolation#3 heliotherapy#1	therapeutic exposure to sunlight
v	2771020	0	0.25	overcloud#1 cloud_up#1 cloud_over#1	become covered with clouds; "The sky clouded over"
v	2771169	0.125	0	light_up#3 clear_up#4 clear#3 brighten#2	become clear; "The sky cleared after the storm"
v	2771320	0	0.25	plague#1 blight#1	cause to suffer a blight; "Too much rain may blight the garden with mold"

Fonte: Autoria própria

Souza, Pereira e Dalip (2017), faz uma comparação entre diferentes abordagens de métodos para a análise de sentimentos usando o *iFeel* (http://blackbird.dcc.ufmg.br:1210/), uma aplicação *web* gratuita que permite detectar opiniões em qualquer texto, usando 18 métodos diferentes. A Figura 3 mostra o gráfico de comparação gerado.

Figura 3 – Gráfico de comparação de métodos em porcentagem de acertos



Fonte: Souza, Pereira e Dalip (2017)

Dentre os métodos usados, o *SentiWordNet* se saiu com maior quantidade de acertos, mesmo que bem próximo a outros. Vale notar que os dois primeiros léxicos, o *UniLex1* e o *UniLex2*, foram os desenvolvidos por Souza, Pereira e Dalip (2017) e não foram encontrados disponíveis publicamente para uso.

CRIAÇÃO DO LÉXICO

Para a elaboração do léxico usado neste projeto foi utilizado o método de expansão a partir da conjugação de verbos. O dicionário escolhido como base foi o *SentiWordNet*. A seguir é apresentado o tratamento realizado no *SentiWordNet* e as adições feitas ao léxico.

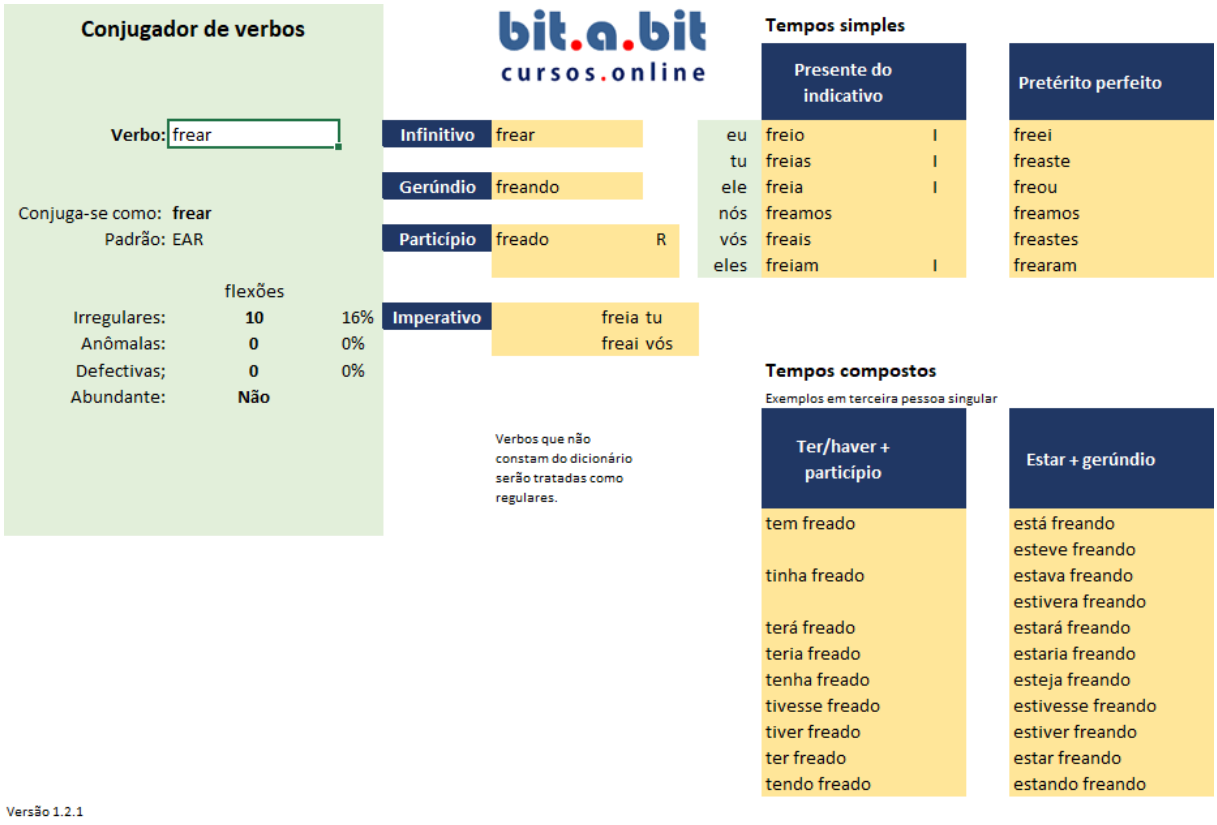
Usando o pacote lexicon no R, foi possível ter acesso direto ao dicionário para o uso no projeto. A versão oferecida no pacote possui uma tabela com duas colunas, uma de palavras e expressões e a outra dos valores de positividade correspondentes. A pontuação nessa versão foi calculada pegando a diferença entre os valores positivos e negativos do léxico original, e todos os resultados com valor zero foram removidos (RINKER, 2018). Este dicionário possui 20.093 palavras valoradas com índices de -0.875 até 0.875.

Com este léxico em mãos, começou-se o tratamento. O léxico foi traduzido para a língua portuguesa no R, usando o pacote *RYandexTranslate*. Após a tradução, foram eliminadas as linhas que possuíam expressões, deixando somente as palavras. Depois foi realizado um arredondamento dos índices para -1, 0 e 1. O corte usado foi de -0.25 e 0.25. Dessa forma, palavras com valor maior ou igual a 0.25 foram arredondadas para 1, palavras com valor menor ou igual a -0.25 foram arredondadas para -1, e o restante teve seu valor arredondado para 0. Por fim, foi feita a remoção das palavras que ficaram com valor 0, já que não vão influenciar na análise. O léxico resultante consistiu de 9.018 palavras valoradas como 1 (positiva) ou -1 (negativa).

Após o tratamento, foi feita a expansão deste léxico. Foram selecionadas todas as palavras da base terminadas em ‘ar’, ‘er’, e ‘ir’, com a intenção de separar os verbos. Após isso, foi elaborado um macro no Excel para conjugar todos os verbos encontrados, usando um conjugador de verbos criado por Radamés Manosso em seu curso EAD de Excel.

A planilha (Figura 4) consegue conjugar mais de 13.000 diferentes verbos em suas 65 flexões, incluindo verbos irregulares (MANOSSO, 2017).

Figura 4 – Planilha de conjugação de verbos no Excel



Fonte: Manosso (2017)

Depois de adicionar os verbos conjugados, o léxico contabilizou 6.024 palavras positivas e 6.927 negativas.

RESULTADOS E DISCUSSÃO

A base usada para testar o léxico contém notícias de 1999 até 2017 dos sites Diário do Grande ABC e Jornal do Brasil, base extraída e tratada em um trabalho de Iniciação Científica PIBIFSP feito pelos autores deste resumo em 2017 (MARQUES; COLNAGO, 2017). O banco possui quase 1,5 milhões de notícias em diferentes categorias. Usando uma tabela de palavras e frequências de termos desta base, foi

encontrado que aparecem efetivamente 2.553 palavras positivas e 2.771 palavras negativas distintas.

Para realizar o cálculo do índice de positividade de uma notícia, foi utilizada a Fórmula (1).

$$i = \frac{(N_{positivas} - N_{negativas}) * 100}{N_{total}} \qquad (1)$$

Onde i é o índice de positividade, $\mathrm{N}positivas$ é o número de palavras positivas, $\mathrm{N}negativas$ é o número de palavras negativas, e $\mathrm{N}total$ é o número total de palavras.

Para validação do léxico, foi realizada uma avaliação de quanto os índices se aproximam do sentido real, positivo e negativo, de cada notícia. Foi feita uma classificação manual de uma amostra de 600 notícias escolhidas aleatoriamente, sendo 300 de cada jornal (Diário do Grande ABC e Jornal do Brasil) e 100 de cada categoria (Nacional, Internacional e Economia). A amostra foi classificada por dois autores do presente trabalho, Colnago, G. R. (analista 1) e Castañeda-Ayarza, J. A. (analista 2), que classificaram cada uma das 600 notícias em -1 (muito negativa), -0.5 (negativa), 0 (neutra), 0.5 (pouco positiva) e 1 (muito positiva). Esta análise é subjetiva e leva em consideração que notícias como tragédias ou mortes por exemplo são muito negativas. Índices econômicos ruins como alta inflação e alto desemprego, assim como violência e corrupção são consideradas negativas. Notícias de previsão do tempo e informações sobre o acontecimento de eventos são neutras. Por outro lado, índices indicando crescimento econômico e emprego, por exemplo, são positivos. Por fim, notícias que representem situações como por exemplo acordos de paz/humanitários e finalização de casos de sequestros são considerados muito positivos.

A Tabela 2 mostra os resultados das correlações por jornal e por categoria. Foi utilizada a medida de correlação de Pearson.

Tabela 2 - Correlação das classificações manuais de notícias com as classificações obtidas a partir do léxico

Classificações	Total	Jornais		DGABC			JB		
		DGABC	JB	NAC	INTER	ECON	NAC	INTER	ECON
Analista 1	27%	34%	19%	21%	55%	21%	21%	25%	1%
Analista 2	25%	33%	18%	31%	45%	17%	17%	29%	1%

Fonte: Autoria própria

O resultado da correlação de ambas as classificações (feitas de formas independentes) ficou parecido. Sendo a maior correlação na categoria internacional do Diário do Grande ABC (55% e 45%), e a menor nas notícias de economia do Jornal do Brasil (1%). Em relação à categoria internacional, foi possível observar que muitas de suas notícias negativas são sobre acidentes, ataques terroristas, assassinatos, guerras, entre outros. Como muitas das palavras que aparecem nessas notícias também são negativas pelo léxico (Ex.: ataque, guerra, feridos, assassinato, conflito, morreram, mortos) fica aparente o porquê do sucesso do léxico em classificá-las. Já com a economia, concluiu-se que a maneira que as palavras são usadas (o seu contexto) é mais importante que quais palavras em si aparecem, o que dificultou na classificação do léxico.

A Tabela 3 mostra a tabela de confusão (*confusion matrix*) da avaliação feita a partir do léxico nesta amostra. Os valores -0.5 foram arredondados para -1 e os valores de 0.5 para 1.

Tabela 3 - Matrizes de confusão entre a classificação manual (linhas) e a classificação do léxico (colunas). As matrizes possuem quantidade de notícias (matriz da esquerda) e percentuais (matriz da direita)

		Previsto		
		-1	0	1
Manual	-1	101	94	52
	0	40	128	68
	1	21	44	52

		Previsto		
		-1	0	1
Manual	-1	17%	16%	9%
	0	7%	21%	11%
	1	4%	7%	9%

Fonte: Autoria própria

Dentre as 600 notícias da amostra 47% foram classificadas corretamente. Esta é a soma dos percentuais da diagonal da matriz de confusão, ou seja, 17% foram classificadas corretamente como negativas, 21% como neutras e 9% como positivas. Os maiores erros da classificação são aqueles que são interpretados com o sentimento oposto do real, ou seja, quando uma notícia positiva é classificada como negativa ou vice-versa. A matriz mostra que apenas 13% das notícias foram classificadas dessa forma.

Para o treinamento da rede neural foi usada essa mesma amostra de 600 notícias já classificada. Para avaliação da rede, como não podemos usar a mesma amostra, já que o treinamento foi feito em cima dela, foi classificada uma nova amostra com o mesmo tamanho e proporções da primeira. Esta amostra foi classificada pelo analista 1.

Os resultados da rede são dados em porcentagem de “certeza”, como a rede foi treinada para realizar uma classificação binária (positiva ou negativa), a ela fornece duas porcentagens, uma referindo a categoria positiva e outra a negativa. A soma dessas duas porcentagens sempre iguala a 100%, logo só precisamos de uma para fazer as comparações. Para calcular o índice a partir da porcentagem dada pela rede foi utilizada a fórmula (2).

$$i = P_{positiva} * 2 - 100 \qquad (2)$$

Onde i é o índice de positividade e $Ppositiva$ é a porcentagem positiva dada pela rede neural. Dessa forma, os índices de positividade variam de -1 a 1 (-100% a 100%), onde notícias mais próximas de -1 são negativas, notícias próximas de 1 são positivas e notícias próximas a 0 vão ser consideradas neutras.

A Tabela 4 mostra os resultados das correlações de Pearson dos índices obtidos a partir da rede com a amostra, comparando também esses resultados com as correlações dos índices do léxico com esta nova amostra.

Tabela 4 - Correlação das classificações manuais de notícias com as classificações obtidas a partir do léxico e da rede neural

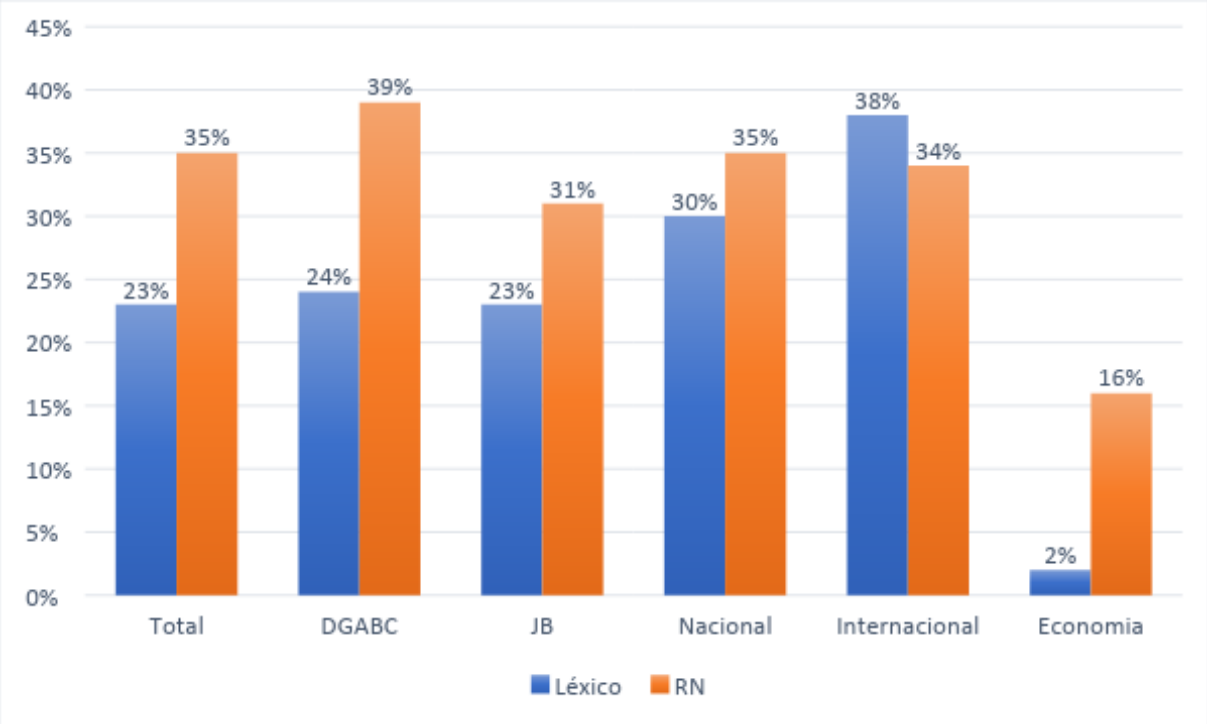
Classificações	Total	Jornais		DGABC			JB		
		DGABC	JB	NAC	INTER	ECON	NAC	INTER	ECON
Rede Neural	35%	39%	31%	38%	36%	10%	28%	32%	22%
Léxico	23%	24%	23%	22%	36%	6%	38%	41%	-1%

Fonte: Autoria própria

As correlações do léxico nesta amostra seguem um padrão parecido com o anterior, a categoria internacional ficando com as maiores correlações (36% no Jornal do Brasil e 41% no Diário do Grande ABC), enquanto economia fica com as menores, (6% e -1% em cada jornal).

Em relação a rede neural, ela classificou de maneira geral melhor que o léxico, tendo um total de correlação maior (35% da rede e 23% do léxico). Agrupando as categorias (Figura 5), é possível ver que a rede neural só possui menor correlação na categoria internacional.

Figura 5 – Gráfico comparativo entre as correlações do léxico e da rede neural com a classificação manual



Fonte: Autoria própria

Além de apresentar melhores resultados de uma forma geral, a rede neural conseguiu ser mais consistente, mesmo na categoria de menores correlações (16% em economia) em teve uma queda muito menor que a do léxico. Como mencionado, a categoria Economia é mais sensível ao contexto em que as palavras estão inseridas. A título de exemplo, as Tabelas 5 e 6 comparam notícias em que a discrepância de classificações ficou mais aparente. Nelas estão apresentadas as notícias, o índice do léxico junto com as palavras consideradas positivas negativas (sem repetição), o índice da rede neural e a classificação manual caracterizada pelos autores.

Tabela 5 – Comparação de notícias da categoria economia com grande discrepância de índices

Notícia do DGABC do dia 18/12/2015

Título: Indústria da construção volta a exibir queda de atividade em novembro, diz CNI

Corpo: Os maiores geradores de riqueza naquele ano foram São Paulo, Rio de Janeiro, Brasília, Belo Horizonte, Curitiba, Manaus e Campos dos Goytacazes. Quando somados os 62 municípios brasileiros mais ricos, chegava-se a metade do PIB nacional. Por outro lado, os 1.388 municípios mais pobres responderam por aproximadamente 1,0% do PIB nacional. Nesse grupo com menor participação na geração de riqueza estavam 74,6% dos municípios do Piauí, 60,1% dos municípios da Paraíba, 53,3% dos municípios do Rio Grande do Norte e 52,5% dos municípios do Tocantins. Segundo o IBGE, não houve alteração significativa entre os municípios com maior participação no PIB no período de abrangência do levantamento, de 2010 a 2013. O município de São Paulo teve o maior recuo em participação na geração de riqueza no País em 2013, segundo a pesquisa do IBGE. A queda foi de 0,4 ponto porcentual em relação a 2012, passando de uma fatia de 11,1% para 10,7% no período. De acordo com o IBGE, a perda foi provocada pelos serviços financeiros, indústria de transformação e comércio de automóveis. Em quatro anos, a participação de São Paulo no PIB nacional recuou 0,8 ponto porcentual. Em 2010, a fatia do município mais rico do Brasil no PIB era de 11,5%. Na direção oposta, o município que apresentou maior aumento em participação no PIB do País na passagem de 2012 para 2013 foi o Rio de Janeiro, que avançou 0,1 ponto porcentual, devido às grandes obras de infraestrutura. As 27 capitais brasileiras responderam juntas por 32,8% da economia do País em 2013, uma redução em relação ao resultado de anos anteriores. Em 2010, as capitais participavam com 34,3%; em 2011, com 33,7%; e em 2012, com 33,4%. Enquanto o município de São Paulo manteve a liderança do ranking de maior geração de riqueza em 2013, Palmas (TO) ocupou o último lugar. Florianópolis (SC) foi a única capital que não tinha o maior PIB entre os municípios de seu estado, onde foi ultrapassada por Joinville e Itajaí. Em 2013, o município de Presidente Kennedy, no Espírito Santo, registrou o maior PIB per capita do País: R$ 715.193,70. Para efeito de comparação, o PIB per capita do País no mesmo ano foi de R$ 26.444,63. No segundo lugar do ranking de maior PIB per capita ficou São Gonçalo do Rio Abaixo, em Minas Gerais, com R$ 340.688,49. Em terceiro, Louveira, em São Paulo, com R$ 278.145,26. Os demais destaques foram: Porto Real (RJ), R$ 255.658,30; Selvíria (MS), R$ 254.242,69; Ilha Comprida (SP), R$ 242.646,02; Quissamã (RJ), R$ 223.042,26; Triunfo (RS), R$ 215.393,60; São João da Barra (RJ), R$ 212.966,61; e Itapemirim (ES), R$ 187.712,94. O IBGE informou que Ilha Comprida (SP), Quissamã (RJ), São João da Barra (RJ) e Itapemirim (ES) eram produtores de petróleo. Em São Gonçalo do Rio Abaixo (MG), a extração de minério de ferro é a principal atividade econômica. Louveira (SP) era sede de centros de distribuição. Porto Real (RJ) sedia uma indústria

automobilística. Selvíria (MS) produzia eucalipto para as indústrias de celulose e possuía hidrelétrica. Triunfo (RS) era sede de um polo petroquímico importante. O menor PIB per capita do País em 2013 foi do município de Nina Rodrigues, no Maranhão, apenas R$ 3.241,29. A economia local sustentava-se pela transferência de recursos federais: 63,4% do valor adicionado bruto total do município vinham da Administração Pública. Entre as capitais, Vitória (ES) tinha o maior PIB per capita em 2013, R$ 64.001,91, enquanto Maceió (AL) tinha o menor, R$ 16.439,48. Apesar dos avanços registrados nos últimos anos na redução da desigualdade, a riqueza permanece bastante concentrada no País. Em 2013, apenas sete municípios concentravam 25% da economia brasileira, de acordo com o Produto Interno Bruto dos Municípios 2010-2013, divulgado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) nesta sexta-feira, 18.

Índice do Léxico: 3,05 (positiva)
Palavras positivas do léxico (24): maior riqueza abrangência obras resultado triunfo extra principais valor administração vitória
Palavras negativas do léxico (12): recuo provocou econômica palmas efeito comparação bruto

Índice da RN: -0,1746 (negativa)

Classificação manual: negativa

Fonte: Autoria própria

Tabela 6 – Comparação de notícias da categoria economia com grande discrepância de índices

Notícia do DGABC do dia 19/08/2010

Título: Volks abre 312 vagas para fábrica de São Bernardo

Corpo: A Volkswagen do Brasil está contratando 312 funcionários para sua fábrica Anchieta, em São Bernardo, "como forma de ampliar sua capacidade de produção", diz a montadora, em comunicado. Os novos colaboradores atuarão nas áreas de armação, pintura e montagem final. "Essas contratações demonstram nossa confiança no mercado brasileiro e representam mais um passo em nossa estratégia de crescimento sustentado, que tem como meta vender 1 milhão de unidades em 2014, quando o mercado total brasileiro deverá chegar a 4 milhões de unidades", afirmou o presidente da Volkswagen do Brasil, Thomas Schmall, no comunicado. No fim do ano, a montadora anunciou investimentos da ordem de R$ 6,2 bilhões no País, dinheiro que será desembolsado entre 2010 e 2014. Com os aportes previstos para este ano e as atuais contratações, a capacidade de produção da fábrica Anchieta será elevada dos atuais 1.300 veículos por dia para 1.600 unidades diárias. Os recursos estão sendo aplicados no desenvolvimento de produtos e na expansão da capacidade produtiva da fábrica de veículos de Taubaté e da unidade de motores, em São Carlos. A companhia vai apresentar ao mercado em 2010 13 lançamentos. O Sindicato dos Metalúrgicos do ABC destaca a atuação da entidade para garantir mais postos de trabalho e revela que a montadora contratou 3.609 trabalhadores, somente para a planta Anchieta, desde 2007. Desses, 3.101 foram efetivados.

Índice do Léxico: -3,08 (negativa)
Palavras positivas do léxico (6): forma ampliar pintura demonstra confiança expansão
Palavras negativas do léxico (10): contrato fabrica final crescimento vai

Índice da RN: 0,4132 (positiva)

Classificação manual: positiva

Fonte: Autoria própria

CONCLUSÕES

Léxicos são recursos muito utilizados no Processamento de Linguagem Natural e, por extensão, na Mineração de Textos, inclusive na análise de sentimento. Porém, existem muitas outras técnicas que envolvem este tipo de análise, como a utilização de aprendizagem de máquinas em redes neurais artificiais. Com isto, este trabalho se foca na criação e avaliação de um léxico de sentimentos para a língua portuguesa, no treinamento e avaliação de uma rede neural recorrente com LSTM e na análise comparativa dos resultados obtidos com as duas técnicas. O léxico atuou melhor em situações onde a distinção entre positivo/negativo era mais aparente, ou seja, baseada em termos isolados, mas falhou em situações que necessitavam de mais contexto. Já a rede neural apresentou no geral melhores resultados que o léxico, conseguindo também ser mais consistente em suas previsões. Este resultados evidenciam a complexidade da linguagem, mostrando a importância de considerar o contexto em que as palavras estão inseridas, o que é melhor captado com as redes neurais. Este trabalho faz parte de um Projeto de Iniciação Científica em desenvolvimento pelos autores. A análise de sentimentos de notícias possui um grande potencial, podendo ser aplicadas análises como a análise de tendência de positividade por jornal/categoria ao longo do tempo, a análise de tendência de diferentes jornais com relação a determinados assuntos, a relação de cotações de empresas com os índices de notícias sobre as mesmas, análise de sentimento de notícias em diferentes segmentos de empresas, entre outras. Estes são alguns exemplos de propostas a serem desenvolvidos pelos autores.

AGRADECIMENTOS

Agradecemos ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Câmpus Cubatão e ao Programa Institucional de Bolsas do IFSP (PIBIFSP).

REFERÊNCIAS

DODDS, P. S. et al. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. **PLoS ONE**, 2011.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de informação. **Ci. Inf. [online]**, 2006.

FREITAS, C. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **Revista Brasileira de Linguística Aplicada**, 2014.

GODBOLE, N.; SRINIVASAIAH, M.; SKIENA, S. Large-Scale Sentiment Analysis for News and Blogs. IN:INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA. **Proceedings of…** [S.l.]: [s.n.]. 2007.

LABILLE, K.; GAUCH, S.; ALFARHOOD, S. **Creating Domain-Specific Sentiment Lexicons via Text Mining**, 2017.

MANOSSO, R. **Conjugador de verbos em Excel**. [S.l.]. 2017.

MARQUES, G. M.; COLNAGO, G. R. **Text Mining na Classificação de Notícias**. [S.l.]. 2017.

MICROSOFT. **SQL Server 2016 SP1 Express Edition**. Disponivel em: <https://www.microsoft.com/en-us/sql-server/sql-server-editions-express>. Acesso em: 25 Junho 2017.

PANG, B.; LEE, L. **Opinion Mining and Sentiment Analysis**. Boston: Now Publishers Inc., 2008.

PYTHON. Disponivel em: <https://www.python.org>. Acesso em: 25 Junho 2017.

RINKER, T. **Lexicons for Text Analysis**. [S.l.]. 2018.

ROLIM, V.; FERREIRA, R.; COSTA, E. D. B. Utilização de Técnicas de Aprendizado de Máquina para Acompanhamento de Fóruns Educacionais. **Revista Brasileira de Informática na Educação – RBIE**, 2017.

RSTUDIO TEAM. **RStudio**: Integrated Development Environment for R, 2016. Disponivel em: <https://www.rstudio.com>. Acesso em: 25 Junho 2017.

SOUZA, K.; PEREIRA, M.; DALIP, D. H. **UniLex**: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro, 2017.

THE R FOUNDATION. **The R Project for Statistical Computing**. 2017. Disponivel em: <https://www.r-project.org>. Acesso em: 25 Junho 2017.