Construção de Dicionário de Palavras para Análise de Sentimentos em Tweets

Lígia Iunes Venturott[1], **Patrick Marques Ciarelli**[2]

[1]Departamento de Engenharia Elétrica – Universidade Federal do Espírito Santo (UFES)
Vitória – ES – Brazil

ligia.venturott@gmail.com, patrick.ciarelli@ufes.br

Abstract. *The recent growth of social networks provided us with a great amount of information about their users, which can be of interest to some sectors of society, such as governments or companies. A useful analysis of this data is identifying the sentiment polarity in texts about some specific subject. Many computer methods have been proposed, but due to their complexity they are not simple to understand, which is not ideal in some scenarios. In this paper we develop an approach of sentiment analysis of tweets that can be easily understood and adapted by others. The approach consists in creating a dictionary of words that gives each word a value according to its polarity. To estimate the weights of each word a Genetic Algorithm was used. The resultant dictionary allows for easy interpretation and modification. When compared to other methods, the results show a promising approach.*

Resumo. *O crescimento das redes sociais proporcionou uma grande disponibilidade de informações sobre seus usuários, que são de importância para um grande número de interessados, como setores do governo ou empresas. Uma análise útil desses dados é identificar a polaridade do sentimento em um texto sobre determinado assunto. Muitos métodos computacionais já foram propostos neste sentido, porém, devido à sua complexidade , eles costumam ser de difícil interpretação, o que pode não ser ideal para determinados cenários. Neste trabalho buscou-se desenvolver um método de análise de sentimentos de tweets que seja de fácil modificação e fácil interpretação para usuários leigos. A abordagem desenvolvida consiste em criar um dicionário de palavras que pondera cada palavra de acordo com a sua polaridade. Para estimar os pesos de cada palavra foi utilizado um Algoritmo Genético. O dicionário formado permite fácil interpretação e modificação. Quando comparado com outros métodos, os resultados indicaram que a abordagem proposta é promissora.*

1. Introdução

Na última década, as redes sociais passaram por uma grande expansão e aumento da popularidade. Com isso houve um grande aumento na quantidade de textos publicados. Apenas na rede social Twitter foi registrada, em 2018, uma média de 500 milhões de publicações por dia [Aslam 2018].

As redes sociais permitiram que qualquer indivíduo pudesse expressar seus pensamentos para o resto do mundo, e suas publicações contêm as opiniões, gostos, preferências e hábitos de cada usuário. Essas características as fazem fontes valiosas de

dados, pois podem prover informações como o nível de satisfação de consumidores sobre um determinado produto, ou a opinião geral da população sobre um determinado governo, por exemplo, tornando-as de grande interesse para vários setores da sociedade. Além disso essa grande massa de dados pode ser utilizada também por pesquisadores. No entanto, devido ao seu grande volume, a análise manual dos dados torna-se impraticável, sendo necessário substituir o o trabalho manual pelas máquinas.

Um tipo específico de análise que pode ser feita sobre esses dados é a análise de sentimentos. Ela busca extrair dos textos o sentimento atribuído pelo autor. Quando aplicada sobre dados de uma rede social pode ser útil para descobrir o grau de aceitação das pessoas em relação a um determinado assunto.

Atualmente existem diversos métodos na literatura que realizam análise de sentimentos, porém boa parte é do tipo caixa preta, pois não é claro como os dados são usados para se chegar nos resultados esperados, sendo as redes neurais um exemplo deste tipo de abordagem. Isso faz com que os resultados desses métodos sejam de difícil interpretação, principalmente para pessoas que não dominam o seu entendimento.

Dentre os métodos que não se encaixam no tipo caixa preta é importante citar o Emolex [Araújo et al. 2014], que serviu de inspiração para esse trabalho. O método tem o objetivo de não só determinar a polaridade de frases, mas também os sentimentos predominantes nela. Para isso, ele utiliza um dicionário de palavras, cada uma associada a valores de 10 diferentes parâmetros, 8 emoções básicas propostas por [Plutchik 1982], e as polaridades positivo e negativo. Esse dicionário de palavras e suas associações com as emoções foram construídas manualmente por uma equipe.

Sendo assim, este trabalho tem como foco desenvolver um método de análise de sentimentos que seja de fácil interpretação dos resultados, e de fácil adaptação por pessoas leigas. A proposta é construir um dicionário de palavras contendo suas polaridades e intensidades, porém, diferentemente do Emolex, as atribuições dos pesos das palavras será por meio de um algoritmo Genético, de maneira que qualquer pessoa possa modificar o dicionário de maneira fácil, e que haja um entendimento do funcionamento do método.

2. Fundamentação Teórica

2.1. Análise de Sentimentos

Análise de Sentimentos é a área que estuda a extração de informação subjetiva de textos, os sentimentos [Pang et al. 2008]. O principal desafio dessa tarefa é a grande complexidade que apresentam as linguagens naturais. Elas são naturalmente entendida por nós humanos, porém apresentam uma estrutura sem regras tão rígidas e sempre em mutação, o que as torna, em alguns aspectos, até ilógicas.

Algumas características presentes em todas as línguas, como a ambiguidade de palavras ou as múltiplas maneiras de expressar o mesmo sentido, tornam a linguagem natural um objeto de difícil interpretação por uma máquina.

Existem diferentes técnicas de Análise de Sentimentos que utilizam diferentes áreas da inteligência artificial. Porém, um fator comum entre elas é converter as características das palavras em forma numérica para calcular o sentimento ou polaridade dos textos. Nessas técnicas a definição dos valores numéricos e como manipulá-los para obter o resultado esperado são feitos utilizando algum método de inteligência artificial.

2.2. Algoritmo Genético

O algoritmo genético é um algoritmo de otimização inspirado na teoria da evolução das espécies de [Darwin 1859]. O algoritmo simula o processo de seleção natural utilizando também fatores como mistura genética e mutação para garantir a variabilidade das soluções.

Nesse algoritmo as possíveis soluções para o problema são denominadas indivíduos, e um conjunto dessas soluções, população. O processo se inicia com a criação, aleatória ou não, de uma população inicial. É definida uma função chamada *fitness function* que determina quão boa é cada solução. É nessa função que é explicitado o problema a ser otimizado.

Após o cálculo da *fitness* de cada indivíduo selecionam-se os melhores, ou mais adaptados, como no processo de seleção natural. É feita então uma mistura das características dos indivíduos selecionados para gerar novos indivíduos, simulando a reprodução e mistura de genes. Por fim algumas mudanças aleatórias são feitas na população resultante, simulando a mutação. Com a população final, repete-se o processo por uma quantidade de ciclos, gerações, pré-determinados ou até que algum membro da população consiga uma *fitness* boa o suficiente.

O algoritmo genético se destaca entre os métodos de otimização por obter resultados satisfatórios em um período muito menor de tempo do que a busca exaustiva [Holland 1973].

2.3. Aprendizado de Máquina

Aprendizado de máquina é o ramo da inteligência artificial que é usado para extrair informações de conjuntos de dados. Tais informações são expressas de uma forma compreensível e podem ser usadas para uma variedade de propósitos [Witten et al. 2016].

É uma área que cresceu muito recentemente devido ao grande volume e variedade de dados disponíveis. Esse ramo possui diversos subramos com inúmeros métodos de aprendizagem desenvolvidos. A aprendizagem se baseia em minimizar o número de erros cometidos pela máquina, ou maximizar o número de acertos, o que no final se resume muitas vezes a um problema de otimização. Isso abre para a possibilidade de usar algoritmos de otimização como o algoritmo genético.

2.4. Processamento de Linguagens Naturais

Linguagem natural é o termo utilizado para designar os métodos de comunicação utilizados no dia-a-dia por seres humanos, como as línguas inglesa, francesa e portuguesa. Processamento de Linguagem Natural, *Natural Language Processing - NLP* em inglês, é a área da inteligência artificial dedicada a fazer com que computadores entendam e processem a linguagem natural, tanto para melhorar as interações homem-máquina como para extrair informações de linguagens naturais [Brownlee 2017].

Alguns desafios de NLP a serem citados podem ser: reconhecimento de fala, tradução instantânea, geração de textos e *chatbots*.

3. Metodologia

Nessa seção são descritas as metodologias aplicadas neste trabalho como também os materiais utilizados.

Como base de dados para esse projeto foram coletados em 2017 tweets sobre o tema ENEM 2017. A base de dados original contém 2356976 tweets.

Para o pré-processamento dos tweets foi utilizada a ferramenta Cogroo [USP 2011] que devolve a forma canônica das palavras, além de sua classe gramatical. Com isso foram selecionadas apenas as palavras pertencentes às classes gramaticais: substantivo, verbo, adjetivo, advérbio. Foram retiradas também as chamadas *StopWords*, uma lista de palavras extremamente comuns em uma língua ou que não trazem algum significado importante.

Após o tratamento dos tweets o dataset é passado para uma forma vetorial em que cada palavra da frase é representada por seu índice no dicionário de palavras inicialmente sem valores.

Para classificar a base de dados ENEM 2017 foi utilizada a heurística proposta em [Cavalcante e Barbosa 2017], onde se procura dentro das frases emoticons, e a partir desses emoticons se classificam as frases binariamente, como positiva ou negativa. Para essa abordagem foram utilizadas algumas regras:

- São aceitos apenas tweets que contém algum emoticon;
- Não são aceitos tweets que contém tanto emoticons positivos quanto negativos ao mesmo tempo;
- O tweet não pode ser composto apenas por emoticons;
- Não são aceitos retweets.

Os emoticons utilizados para classificação estão expostos na tabela abaixo. Ao final, foram rotulados 10772 tweets (aproximadamente 70% negativos e 30% positivos).

Tabela 1. Classificação de Emoticons

Emoticons Positivos	:)	:-)	:3	=)	:D	xD	XD	=3
Emoticons Negativos	:(	:-(	:/	:\\	:c	:(	=/	=\\

O programa iFeel [Araújo et al. 2014], em Java , contém vários métodos de análise de sentimentos para a língua inglesa, incluindo o Emolex [Mohammad e Turney 2013] que foi utilizado para comparação de resultados.

O método proposto para realizar a análise de sentimentos foi inspirado no Emolex, no qual é utilizado um dicionário de palavras que foram analisadas manualmente e cada uma recebe um vetor de 10 valores binários, 8 para as emoções básicas de [Plutchik 1982] e 2 para as polaridades positivo e negativo. Nessa abordagem foi feita uma simplificação. Ao invés de 8 emoções, foram utilizadas apenas as polaridades positiva e negativa. A determinação do peso das palavras foi feita a partir dos textos previamente rotulados.

Um algoritmo genético (AG) [Houck et al. 1995] implementado em MATLAB foi utilizado para chegar ao peso de cada palavra.

Baseando-se nesse método foram criadas 2 abordagens para serem testadas.

3.1. Classe da Palavra

Nessa abordagem a avaliação da polaridade da frase é feita somando os pesos de cada palavra contida na frase, sendo o valor final da frase dado por:

$$F = \sum_{i=1}^{n} W(P0_i)$$

Em que F é o valor do tweet, $P0_i$ o índice da palavra no dicionário, $W(PO_i)$ o peso da palavra PO_i no dicionário e n a quantidade de palavras no tweet.

As frases com valores finais entre -0,5 e 0,5 são consideradas neutras. Frases com pontuações maiores do que 0,5 são avaliadas como positivas, e abaixo de -0,5 como negativas.

Para definir o peso que cada palavra terá, foi usado um algoritmo genético que busca o peso de cada palavra em um intervalo entre -2 e 2.

3.2. Contexto da Palavra

Nessa abordagem considerou-se o contexto em que cada palavra está inserida na frase para representá-la com uma maior riqueza de informações. Cada palavra é representada por seis parâmetros: o seu índice no dicionário e outras cinco informações extraídas sobre ela.

Tabela 2. Parâmetros das palavras usadas na abordagem contexto da palavra.

P0	Índice da Palavra no Dicionário	1 a ∞
P1	Se é a primeira palavra da frase	0 ou 1
P2	Se foi escrita inteiramente em maiúsculo	0 ou 1
P3	Se possui vogais repetidas (GOOOOL, siiiim)	0 ou 1
P4	Se é uma palavra de negação (não, nunca, etc)	0 ou 1
P5	Distância para palavra de negação que a precede na frase	0 a ∞

Cada palavra da frase é representada nesta abordagem por um vetor de seis parâmetros ao invés de um único número. Abaixo é ilustrado um exemplo de representação de uma frase em vetor com essa abordagem, sendo que a palavra "um" foi removida por ser considerada uma *StopWord*.

Hoje não é um bom dia.$\rightarrow [110000; 200010; 300001; 400002; 500003]$

Utilizando o algoritmo genético, a cada palavra do dicionário é atribuída um peso de -2 a 2, e para os parâmetros de contexto são atribuídos pesos entre 0 e 2.

Foram criados 2 parâmetros α e β para ajustar a importância do parâmetro P5 no cálculo do valor da frase. Os dois parâmetros também são aprendidos pelo AG, de forma a melhor aproveitar o valor de P5. O resultado é um dicionário de $m + 7$ pesos, sendo m o número de palavras no dataset e 7 o número de parâmetros de contexto mais os pesos extras α e β.O cálculo do peso da palavra se deu por 2 fórmulas:

- P5 > 0 : há uma palavra negativa na frase, porém não é a palavra avaliada:

$$V_i = W(P0_i) * (P1_i * W_{P1} + P2_i * W_{P2} + P3_i * W_{P3}) * tanh(\alpha * P5_i + \beta))$$

- P5 = 0: não há palavra negativa na frase, ou a palavra negativa é a avaliada:

$$V_i = W(P0_i) * (P1_i * W_{P1} + P2_i * W_{P2} + P3_i * W_{P3}) * 2 * (0,5 - P4_i))$$

O cálculo do valor do tweet é a soma dos valores de cada palavra:

$$F = \sum_{i=1}^{n} V_i$$

4. Experimentos e Resultados

Nesta seção serão apresentados a base de dados usada nos experimentos, a configuração dos experimentos e os resultados.

Foi utilizada para os experimentos a base de dados ENEM 2017, com os tweets classificados pela heurística proposta em [Cavalcante e Barbosa 2017] Para os experimento, os tweets foram divididos de maneira aleatória, 50% para treino e 50% para teste, porém de forma que os conjuntos de teste e treino possuíssem a mesma proporção de tweets positivos/negativos.

Foram feitos experimentos com 4 abordagens: os dois métodos propostos (frequência de palavras e contexto da palavra), Emolex traduzido e a rede LSTM (*Long short-term memory*) [Gers et al. 1999].

O método Emolex, utilizado por meio do programa iFeel, foi usado como inspiração para os métodos propostos neste trabalho, e por isso tais métodos serão comparados com uma versão traduzida do seu dicionário realizada pelos autores deste trabalho. Porém como o método foi criado para a língua inglesa, seu dicionário foi traduzido para a língua portuguesa. O seu modo de funcionamento é semelhante ao do método Classe da Palavra, com a diferença que os pesos das palavras foram obtidos por análise de especialistas.

A rede neural utilizada possui uma camada de *embedding* pré-treinada, que utilizou um *embedding* de 100 dimensões descrito em [Hartmann et al. 2017]. A camada de *embedding* é seguida por uma camada LSTM contendo 15 células e uma *Dense Layer* como camada de saída. A implementação foi feita utilizando a API Keras para Python.

A acurácia foi calculada de maneira simples:

$$Acc = \frac{\#\ Tweets\ \ classificados\ \ corretamente}{\#\ Tweets\ \ Totais}$$

Os resultados dos experimentos estão apresentados na tabela abaixo.

Tabela 3. Resultados dos Experimentos

Classe da Palavra	Contexto da Palavra	Emolex Traduzido	LSTM
74,5%	74,5%	74,1%	75,0%

Os quatro métodos testados apresentaram resultados muito próximos. Apesar de ter uma maior complexidade o método Contexto da Palavra apresentou o mesmo desempenho do método Classe da Palavra.

Os dois métodos desenvolvidos apresentaram desempenho superior ao método do Emolex traduzido, apesar de pequena a diferença no resultado. Porém, os pesos do dicionário do Emolex foram dados utilizando como base textos provindos de fontes diferentes do Twitter, com outros contextos e em outra língua. Já nos métodos desenvolvidos, os pesos das palavras são definidos pelo Algorítimo Genético baseando-se na própria base de dados ENEM 2017 em que eles foram testados, deixando o dicionário criado mais adaptado às características textuais de um ambiente como o Twitter.

Já a rede neural alcançou um desempenho maior do que os outros métodos, o que já era esperado por ter sido construída para lidar com dados sequências, como por exemplo textos.

5. Conclusões

Este trabalho teve como objetivo desenvolver um método para análise de sentimentos de fácil entendimento e utilização, utilizando o algoritmo genético.

Como base de dados foram utilizados textos extraídos do Twitter. Foi perceptível durante a realização do trabalho as dificuldades impostas pelo tipo de linguagem que é utilizada nesse tipo de rede social. Por se tratar de uma linguagem extremamente informal com presença de muitos erros ortográficos, sejam eles propositais ou não, emojis e referências a outros usuários, o texto torna-se difícil de tratar, sendo necessária uma ampla etapa de pré-processamento. Outra dificuldade encontrada é a repetição de textos na amostras, os retweets, que devem ser removidos antes do processamento.

Pela comparação dos resultados com o método Emolex e com a rede neural pode-se perceber que o método desenvolvido é consistente. Em trabalhos futuros é possível utilizar melhores técnicas de pré processamento para tentar amenizar o problema causado pela procedência do texto. Uma outra abordagem possível seria considerar que palavras são bastante complexas para expressar com apenas um peso, e que mais pesos poderiam ser atribuídos a cada palavra. Isso porém aumentaria o tempo de processamento para a geração do dicionário.

Referências

Araújo, M., Gonçalves, P., Cha, M., e Benevenuto, F. (2014). ifeel: a system that compares and combines sentiment analysis methods. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 75–78. ACM.

Aslam, S. (2018). Omnicore. https://www.omnicoreagency.com/twitter-statistics/. Acessado em 11/05/2018.

Brownlee, J. (2017). What is natural language processing? https://machinelearningmastery.com/natural-language-processing/. Acessado em 20/10/2018.

Cavalcante, P. E. C. e Barbosa, Y. d. A. M. (2017). Um dataset para análise de sentimentos na língua portuguesa.

Darwin, C. (1859). *On the origin of species, 1859*.

Gers, F. A., Schmidhuber, J., e Cummins, F. (1999). Learning to forget: Continual prediction with lstm.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., e Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Holland, J. H. (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing*, 2(2):88–105.

Houck, C. R., Joines, J., e Kay, M. G. (1995). A genetic algorithm for function optimization: a matlab implementation. *Ncsu-ie tr*, 95(09):1–10.

Mohammad, S. M. e Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Plutchik, R. (1982). A psychoevolutionary theory of emotions.

USP (2011). Cogroo - corretor gramatical acoplável ao libreoffice. http://cogroo.sourceforge.net/index.html. Acessado em Março de 2018.

Witten, I. H., Frank, E., Hall, M. A., e Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.