UNIVERSIDADE FEDERAL DE LAVRAS

Métodos para a Análise de Sentimentos

GCC151 - 3ª Avaliação

Alunos: Franciscone Almeida, Igor Cruz e Victor Hugo Landin.

LAVRAS - MG
2019

Introdução

Atualmente, com o advento da internet e das mídias sociais e com o avanço crescente da tecnologia de informação, há um grande aumento da conectividade entre pessoas do mundo todo e é cada vez mais comum que grandes empresas façam uso de métodos de inteligência artificial para melhorar cada vez mais seus serviços. A base de informações que é gerado na internet vem se tornando imensurável e extrair conhecimento a partir de dados desse tipo se tornou muito importante para diversas áreas como entretenimento, comércio, opinião pública etc.

Dentre tais métodos, encontra-se a Análise de Sentimentos, no campo do Processamento de Línguas Naturais, que é caracteriza por identificar emoções em opiniões expressas em determinados textos, classificando-as como positivas, negativas ou até neutras. Esta tarefa é muito utilizada para extrair polaridades em comentários e avaliações de entidades, sendo muito importante para obter informações sobre o grau satisfação dos usuários em relação à itens e também informações mais gerais sobre pontos fortes e fracos, auxiliando as empresas na coleta de informações sobre seus produtos e/ou prestação de serviços.

Referencial Teórico

Na Análise de Sentimentos existem diversas técnicas para fazer a análise dos dados, tais como: Emoticons, LIWC, SenticNet e Happiness Index. Cada um desses métodos é mais apropriado e apresenta melhor acurácia para conjuntos de problemas diferentes, os quais serão descritos com mais detalhes nessa seção.

1. Emoticons: O uso de emoticons cresceu juntamente com o uso das redes sociais, e talvez essa seja uma das formas mais simples de se identificar a polaridade em um texto. Os Emoticons são primordialmente baseado em expressões faciais (embora existem algumas exceções como o coração <3) e podem expressar sentimentos de felicidade ou tristeza.
2. LIWC: é uma ferramenta para análise de texto que estima componentes emocionais, cognitivos e estruturais de um dado texto baseada no uso de dicionários contendo palavras e suas respectivas categorias. É uma ferramenta de caráter mais comercial.
3. SenticNet: é um método para mineração de opinião e análise de sentimentos que explora técnicas de Inteligência Artificial. O objetivo do SenticNet é inferir polaridade de textos em nível semântico, e não sintático. A ferramenta foi testada pelos autores para medição de níveis de polaridade em opiniões de pacientes sobre o serviço de saúde nacional na Inglaterra.
4. Happiness Index: é uma escala de sentimentos que utiliza o ANEW. O Happiness Index foi construído com base no ANEW e calcula pontuações com valores entre 1 e 9 que indicam a quantidade de felicidade naquele texto fornecido como entrada para

os testes. No caso do Happiness Index, o valor 1 indicaria o menor índice de felicidade e o valor 9 o maior índice.

Métodos e Resultados

Com base nas técnicas e ferramentas supracitadas, o objetivo desse trabalho era analisar um corpus com opiniões de clientes do buscapé, treinar um algoritmo de aprendizagem de máquina e usar o modelo obtido para classificar a polaridade de uma sentença ou de um texto. Durante o desenvolvimento do trabalho realizamos testes com 4 algoritmos diferentes: Logistic Regression (LR), Multilayer Perceptron (MLP), Random Forest (RF) e o K-Nearest Neighbors (KNN). Comparando os resultados desses algoritmos, todos retornaram uma precisão média de 54%, mas por fim decidimos usar o LR por ser o que apresentava maior acurácia (56,87%).

Nos primeiros testes a precisão do modelo de classificação estava reduzida (em média de 30%) devido a alguns ruídos que haviam no corpus, como as linhas vazias com caracteres especiais no fim (exemplo do ‘\n’ que indica quebra de linha), e ao corrigir esses problemas os resultados foram melhorando.

Percebemos também, que em um certo ponto nenhuma alteração feita, seja no método de treinamento (parâmetros do algoritmo) ou seja nos dados em si, estava fazendo qualquer diferença na melhora da precisão do algoritmo, e isso era devido a dois fatores principais: primeiro, de forma geral, a quantidade de comentarios “inuteis” (ruídos) no córpus (cerca de 3200 arquivos de todo o córpus) que afetavam a integridade dos dados e o modelo de treinamento e segundo a discrepância de valores entre as classes, que nesse contexto era preocupante, pois classes como a 0 e 1 tinham pouco mais de 1000 textos e classes como a 4 e 5 tinham somadas mais de 100.000 textos.

Com base nesses problemas buscamos soluções que pudessem contornar a situação e as duas opções foram: fazer a leitura do arquivo xml do corpus, que continham mais dados como prós e cons de cada produto, aumentando assim os dados e uma segunda abordagem que seria aumentar o corpus atual com mais dados de outros córpus semelhantes.

Ao aplicar as soluções propostas, pode-se observar que no primeiro caso quantidade de dados ainda não era suficiente e os resultados continuaram no mesmo índice (em alguns casos até pioraram). Já na segunda solução, o problema foi encontrar um córpus com a mesma finalidade do fornecido para os testes que estivesse em português. Uma das ideais foi traduzir um córpus do inglês e usá-lo mas a tradução poderia também gerar ruído e afetar na eficácia do algoritmo, além de que o processo de se traduzir cada texto seria custoso e demorado.