Fluxo de Trabalho de um Classificador de Sentimentos

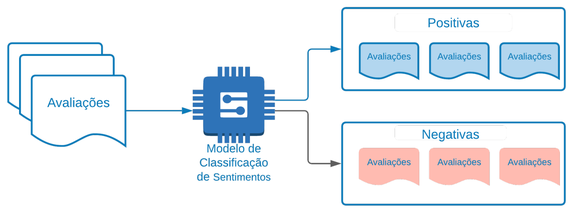
Rafael Morais Rocha
PUC-Minas, Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data
rflmorais@gmail.com

1. Problema Proposto

Todos os dias milhares de avaliações de produtos são feitas pelos consumidores na Internet sobre produto. Saber o se os usuários estão falando bem ou mal do seu produto é algo vital para um vendedor. Por ser um repositório alimentados pelos usuários, seu crescimento é rápido e contínuo. Desta forma, devido a sua grande magnitude, uma análise manual para extrair o sentimento da avaliação gerada pelos usuários é algo impraticável.
Portanto, a classificação automática de conteúdo textual torna-se o único método prático para classificação e percepção eficazes de dados. Assim, o problema proposto é o seguinte: dado uma avaliação textual produzida pelo usuário/consumidor será atribuído, por meio de um algoritmo de aprendizado de máquina, um rótulo ”positivo”ou ”negativo”baseado no conteúdo do texto da avaliação.

2. Saídas/Predições

Neste projeto serão realizadas tarefas baseadas em classificação, um subcampo de Aprendizado de Máquina Supervisionado, onde o objetivo principal é prever rótulos ou respostas de natureza categórica (variável de saída) para os dados de entrada com base no que o modelo aprendeu na fase de treinamento. Os rótulos também são conhecidos como classes, rótulos de classe ou *target (ou label)* são categóricos por natureza, o que significa que são valores não ordenados e discretos. Assim, cada resposta de saída pertence a uma classe ou categoria discreta específica. Assim, nosso problema tem duas variáveis:



- **Variavel** X: é o documento que contém o texto da avaliação, as features.
- **Variável** y: é o rótulo que classifica o texto da avaliação em sentimento positvo (1) ou negativo (0).

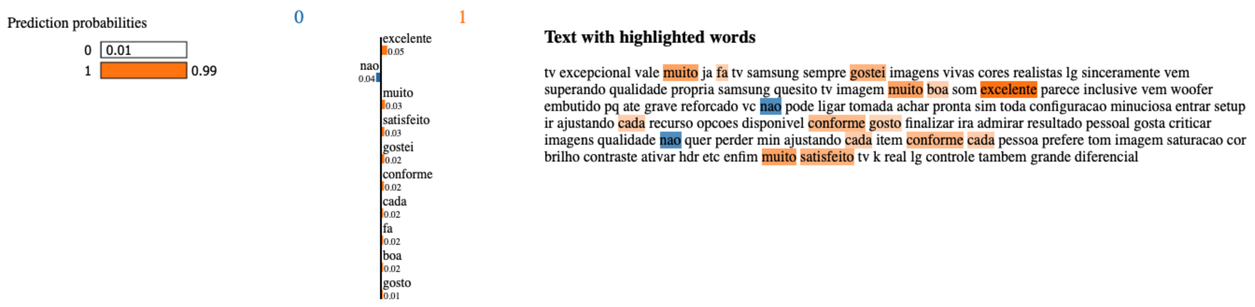
Com estas duas variáveis o modelo será treinado para predizer o rótulo (saída) de novos documentos (entrada).

3. Coleta dos Dados

Os dados para treinar os modelos foram coletados na internet de duas fontes diferentes. Um foi forncecido pela empresa Olist outro pela empresa B2W-Digital. Ambos disponibilizaram o conjunto de dados contento o texto da avaliação é a nota dada pelo avaliador em uma escala de 1 (pior nota) a 5 (melhor nota) estrelas. Os dados estão em um arquivo no formato CSV (*Comma-separeted Values*) e possui documentação expicando os campos. Os dois conjunto de dados, formam um corpus com mais de 200 mil documentos em lingua portuguesa, desta forma o conjunto de dados possui uma boa quantidade de documentos o que nos permite dividir os dataset em dados de treinamento (onde iramos treinar nosso modelo) e dados de teste (onde iremos testar nosso modelo).

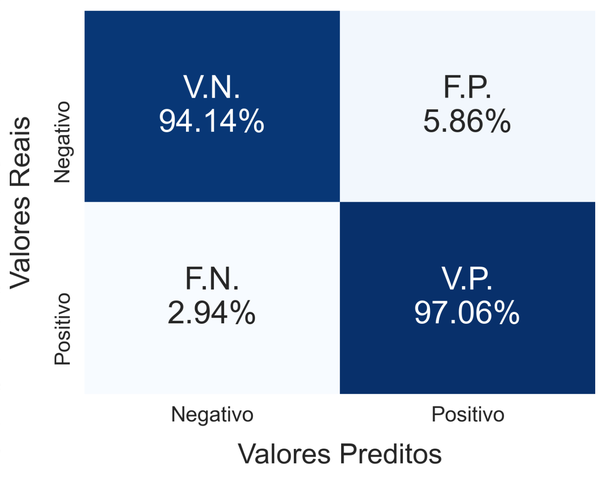
4. Modelagem

As predições do modelo serão categóricas discretas (1 para sentimento positivo, 0 para negativo). Nosso corpus está rotulado, portanto usaremos um um modelo de classificação supervisionado. O modelo irá aprender sobre os dados de treinamento, a partir do qual identificará padrões nas características das palavra do dados de treinamento, para assim atribuir uma classe (ou probabilidade de pertencer à uma classe) a novos documentos. Considerando que são dados de texto, será necessário aplicar os procedimentos de processamento de linguagem natrual e vetorização do texto. Para a modelagem de teste usamos os algoritmos de *Machine Learning*: Multinomial Naïve Bayes, Logistic Regression e Support Vector Machine.
Para criar nosso modelo de classificação de sentimentos, foi escolhido o algoritmo de melhor desempenho, **Logistic Regression**.



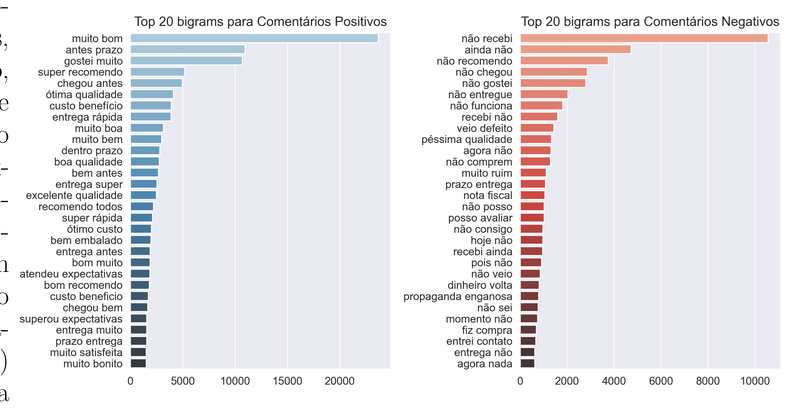
5. Avaliação do Modelo

Para avaliar nosso modelo utlizamos como métricas a acurácia; a precisão; a revocação; a sensibilidade; a media-F ponderada e a área sob a curva ROC (curva AUC).
A acurácia é uma das mediadas mais populares de desempenho, contudo é preciso bastante cuidado em seu uso, principalmente em classificadores binários desbalanceado, já que esta medida não avalia o desempenho de cada classe individualmente. Tendo isso em mente, a principal métrica para comparar e medir o desempenho dos modelos foi a medida F-1 ponderada que permite avaliar como esta sendo o desempenho do modelo por classe. A medida-F é a média harmônica de precisão e revocação. Normalmente é uma medida melhor do que acurácia em conjuntos de dados de classificação binária desequilibrada. O maior valor possível de um *F-score* é 1, indicando precisão e recall perfeitos, e o menor valor possível é 0, se a precisão ou o recall forem zero. **A Matriz de Confusão** apesar de nao ser uma métrica, também possui um papel importante na avaliação do modelo ao summarizar várias delas em um formato tabular.



6. Preparação dos dados

A linguagem humana é complexa e diversa. Os seres humanos se expressam de infinitas maneiras, tanto verbalmente quanto por escrito. Não apenas existem centenas de idiomas e dialetos, como há também um conjunto único de regras gramaticais e de sintaxe, expressões e gírias dentro de cada um deles.
Quando se escreve, é comum cometer erros ou abreviar palavras, ou omitir pontuações. Isso tudo, faz do texto um dado altamente não estruturado. Assim, como estamos trabalhando com textos, precisamos utilizar a metodologia de processamento natural de linguagem, ou seja, além de processar o texto removendo dados inválidos (`NaN`), pontuação, caracteres especiais (RegEx) e *stopwords* é preciso extrair a features do documento de texto em um formato (números) que os algoritmos de aprendizado de máquina possam atuar. Para esta tarefa de *feature extraction* utliziamos da abordagem Term-Frequency (`CountVectorizer`) e Term Frequency-Inverse Document Frequency (`TfidfVectorizer`) e consideraremos uma saco de palavras formados por unigrams e bigrams.



Natural Language Processing Pipeline