Nanodegree - Engenheiro de Machine Learning

Proposta de projeto final

Priscilla Koch Wagner

20 de março de 2019

Histórico do assunto

A disseminação de notícias falsas (fake news em inglês), tem o poder de interferir em diversos setores da sociedade, como política, saúde, segurança, etc. Podem também prejudicar empresas, celebridades, políticos e até pessoas comuns. Qualquer tipo de fake news pode induzir pessoas à erros e ter consequências sérias. Atualmente, por conta da fácil disseminação de informações em redes sociais, as fake news ficaram mais populares. Por isso, a identificação e punição de autores de fake news se tornou uma grande necessidade para autoridades em todo o mundo. Nessa competição para estudantes, disponível na plataforma Kaggle, é disponibilizado um conjunto de notícias com o objetivo de classificar quais são confiáveis e quais não são.

Como irei iniciar um projeto que envolve NLP no meu trabalho, esse desafio é ótimo para testar e aperfeiçoar meus conhecimentos na área. Além disso, tenho me interessado pessoalmente em pesquisar como a disseminação de fake news tem interferido em assuntos importantes e alterando rumo de sociedades ao redor do mundo.

Descrição do problema

O objetivo desse desafio é auxiliar na classificação de notícias confiáveis e não confiáveis através de algoritmos de Machine Learning. Os algoritmos que serão utilizados serão comparados entre si a fim de determinar qual possui maior acurácia para resolver o problema descrito acima. A solução poderá ser comparada com os primeiros vencedores da competição do Kaggle.

Conjuntos de dados e entradas

O conjunto de dados foi obtido a partir da competição para estudantes disponível no Kaggle. A competição, bem como o conjunto de dados pode ser acessado através do link: https://www.kaggle.com/c/fake-news/data.

O conjunto de dados consiste de notícias obtidas durante o último período eleitoral dos Estados Unidos. As notícias foram previamente divididas em conjunto de treinamento e de

teste sendo que existem 20.800 notícias no conjunto de treinamento e 5.200 notícias no conjunto de teste. Cada notícia tem como atributos: um id, o título, o autor e o próprio conteúdo (pode estar inclompleto), e foram classificadas como confiáveis e não confiáveis (rótulo).

Descrição da solução

Dado que o problema envolve análise de texto, a solução utilizará técnicas de NLP (Natural Language Processing). A solução utilizará algoritmos de aprendizagem supervisionada, dado que o conjunto de dados é rotulado. A análise da solução será feita utilizando métricas padrão de análise de algoritmos supervisinados como: acurácia, precision, recall, f1-score.

Modelo de referência (benchmark)

Como benchmark serão consideradas as implementações apresentadas no primeiro kernel do desafio no Kaggle, pelo Paul Larmuseau. Nessa implementação, alguns algoritmos foram sugeridos como possíveis soluções para o problema. Dentre eles, se destacaram: Árvore de Decisão (95% de acurácia) e Regressão Logística (98% de acurácia). O melhor modelo utilizou Regressão Logística regular.

Métricas de avaliação

A principal métrica para avaliação será a acurácia (quanto maior, melhor), pois foi a métrica escolhida para avaliação da competição do kaggle. Os 2 melhores resultados apresentados no leaderboard público da competição do kaggle, possuem uma acurácia de 0.98. Como o objetivo no kaggle é ganhar na pontuação com pequenas mudanças no placar, vamos nos concentrar em obter um modelo com, no mínimo, 0.9 de acurácia, o que me colocaria na posição 8 do placar. O benchmark utilizado, atingiu acurácia de 98% no melhor modelo implementado; e 78% no pior modelo. A acurácia é calculada pela equação abaixo:

$$accuracy = \frac{correctpredictions}{correctpredictions + incorrectpredictions}$$

Além da acurácia, outras métricas serão consideradas (mas não comparadas ao benchmark): precision (quantas notícias preditas como confiáveis são realmente confiáveis), recall(quantas notícias confiáveis foram preditas como confiáveis), f1-score (média harmônica entre precision e recall), curva ROC e KS.

Design do projeto

O projeto será desenvolvido em 6 etapas:

1. Análise exploratória de dados
2. Pré-processamento e feature engineering
3. Treinamento dos modelos
4. Validação e comparação dos resultados dos modelos
5. Entendimento e escolha do melhor modelo

1. Análise exploratória de dados

Nessa etapa, faremos uma análise inicial dos textos das notícias, a fim de obter um melhor esclarescimento sobre os dados e como realizar o pré-processamento na próxima etapa.

2. Feature engineering

Nesta etapa, aplicaremos técnicas de processamento de texto, como Bag of Words ou TF-IDF. Além disso, podemos aplicar PCA após o processamento de texto, a fim de diminuir a quantidade de variáveis.

3. Treinamento dos modelos

Como é um problema de classificação, os algoritmos aplicados serão de aprendizagem supervisionada. Serão aplicados e analisados 4 algoritmos, para encontrar a melhor solução: AdaBoost, RandomForest, SVM com kernel RBF e Regressão Logística com otimização L2.

4. Validação e comparação dos resultados dos modelos

A principal métrica para validação será a acurácia, pois foi a métrica escolhida para avaliação da competição no kaggle. Além da acurácia, outras métricas serão consideradas (mas não comparadas ao benchmark): precision, recall, f1-score, curva ROC e KS. Essas métricas serão utilizadas para comparar os modelos desenvolvidos.

5. Entendimento e escolha do melhor modelo

A seleção do modelo será feita baseada no resultado da comparação das métricas descritas na sessão anterior. O modelo que apresentar a performance que mais se aproxima da solução ideal para o problema será escolhido.