Classificação de sentimento de sentenças utilizando NLP em métodos de aprendizado de máquina

Cristiano M. Matsui[1]

[1]Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

cristiano.matsui@gmail.com

Resumo. *Um dos problemas mais comuns na área de aprendizado de máquinas é a classificação de entidades. Analisar se o objeto retratado em uma imagem é um gato ou um cachorro pode ser trivial para um humano, mas não para um computador. Várias técnicas foram desenvolvidas para buscar solucionar essa família de problemas, dentre elas destaca-se a KNN (k-vizinhos mais próximos) pela sua simplicidade de implementação e a sua eficácia. Este artigo busca solucionar o problema de classificação utilizando base de dados de avaliações de três locais: Amazon, Yelp e IMDB, classificando se o sentimento por trás da avaliação é positivo (satisfação) ou negativo, além de explorar algumas das dificuldades encontradas para lidar com problemas desta família.*

Introdução

O reconhecimento de semântica em falas e sentenças é um conhecido desafio da computação. Embora aconteca de uma maneira extremamente natural e instintiva até mesmo para uma criança, solucionar computacionalmente problemas desta natureza não é uma tarefa trivial. A motivação para atacar este problema engloba interesses como encontrar técnicas que permitam uma máquina a solucionar tarefas práticas, como agendar uma consulta médica ou ligar para uma pizzaria, assim como fornecer modelos que consigam entender o processamento de linguagem natural [Allen 1995].

O tratamento e processamento de uma linguagem natural (falada ou escrita) difere muito do tratamento de uma linguagem estruturada. Esta, foi criada e desenvolvida para propósitos específicos e obedecendo formalismos bem estabelecidos, como linguagens de programação ou a linguagem matemática [Colen 2018]. O processamento de linguagem natural (NLP) envolve interpretação de contexto, sentido e sentimento, promovendo uma interdisciplinariedade de áreas como inteligência artificial, linguística, lógica e psicologia [Joshi 1991].

Este artigo apresenta um problema na área de NLP focado na avaliação e classificação de sentimentos de sentenças retiradas de bases de dados de avaliações escritas por usuários para três plataformas distintas: Amazon, IMDB e Yelp. O interesse principal é elaborar um classificador que consiga distinguir se a sentença escrita possui uma avaliação positiva ou negativa, utilizando ferramentas de aprendizado de máquina como o classificador KNN.

Processamento de Linguagem Natural

Para a utilização das bases de dados de sentenças, algumas etapas de pré-processamento são necessárias. Como a ideia central é a separação por classes (ava-

liação positiva ou avaliação negativa), faz-se necessário extrair características que tornem a classificação possível, assim como retirar possíveis palavras ou símbolos que poderiam atrapalhar o classificador.

Saco de Palavras

O saco de palavras (*bag-of-words*) é um conjunto das palavras presentes no(s) texto(s) de entrada. O interesse neste conjunto não é a ordem em que estas palavras aparecem na sentença, tampouco a gramática por trás destas palavras (tempo verbal, pessoa, gênero), o interesse é puramente na frequência em que estas palavras aparecem. Por isso, é comum aplicar o processo de stemização nessas palavras, antes de tratá-las como tokens. Stemming é o processo de redução de palavras flexionadas ou derivadas para a sua forma raíz, e tem a sua utilidade pois reduz a quantidade de atributos, reduzindo assim a matriz de características[Soares et al. 2009].

Remoção de Palavras Vazias

As palavras vazias (*stop words*) são um conjunto de palavras que geralmente não acrescentam nenhuma características significativa para a etapa de classificação. Essas palavras são conectivos (e, ou), artigos (a, o, um, uma) e demais palavras que não ajudariam o classificador. Esta etapa torna a matriz de características menor e mais significativa. Nesta etapa é vantajoso também remover símbolos e caracteres que não façam parte do alfabeto analizado, como traços, barras e emoticons.

Classificador KNN

O classificador KNN (*k-nearest neighbor*) é um dos classificadores mais simples e mais eficientes encontrados na área de aprendizado de máquina [Tan 2006]. A ideia basica deste classificador é dividir cada elemento da base de dados (neste caso, cada sentença) em um espaço métrico. Utilizando métricas retiradas da análise de características, é possível adicionar novos dados a este espaço métrico. A classificação destes novos dados é feita de acordo com o número k de vizinhos mais próximos a este novo elemento. A sua classificação é então a mesma da maior quantidade de vizinhos mais próximos.

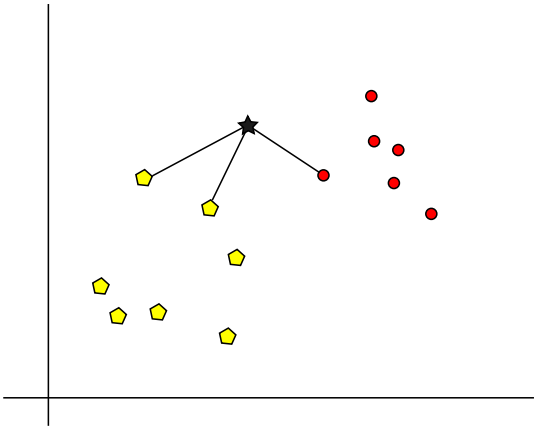

Figura 1. kNN com k=3

Resultados

O principal desafio encontrado foi lidar com a alta dimensionalidade inerente a este tipo de problema. Mesmo após as etapas de stemização e remoção de palavras vazias, foram extraídas 1820 características para a base de dados de avaliações Yelp, 1642 para a Amazon e 2811 para a base da IMDB. Esta dimensionalidade excessiva causou problemas como overfitting do classificador, apresentando taxas de acerto muito alta para o treino e baixas para o conjunto de teste.

Uma solução encontrada para contornar este problema foi reduzir a dimensionalidade utilizando a técnica de Truncated SVD (neste caso de processamento de linguagem natural, Análise Semântica Latente). Com isso, foi possível transformar a matriz esparsa de características em uma matriz menor e mais densa, contornando o problema da dimensionalidade. Abaixo, uma tabela com os resultados encontrados utilizando as matrizes de dimensão original e as reduzidas. Todas etapas foram feitas com um valor fixo de k = 3, afim de medir os efeitos da redução da dimensionalidade no classificador. Os resultados foram validados através da técnica de validação cruzada, utilizando 10% da base de dados como holdout.

Base de dados	**Taxa (treino)**	**Taxa (teste)**
Amazon	77,9%	63,1%
Yelp	93,3%	67,8%
IMDB	77,6%	62,6%

Tabela 1. Taxa de acerto do classificador - matriz original

Base de dados	**Taxa (treino)**	**Taxa (teste)**
Amazon	81,8%	80,3%
Yelp	82,9%	81,0%
IMDB	79,4%	77,8%

Tabela 2. Taxa de acerto do classificador - matriz reduzida

Base de dados	**Matriz (treino)**	**Matriz (teste)**
Amazon	$\begin{bmatrix} 309 & 191 \\ 30 & 470 \end{bmatrix}$	$\begin{bmatrix} 444 & 78 \\ 128 & 350 \end{bmatrix}$
Yelp	$\begin{bmatrix} 472 & 28 \\ 39 & 461 \end{bmatrix}$	$\begin{bmatrix} 410 & 93 \\ 112 & 385 \end{bmatrix}$
IMDB	$\begin{bmatrix} 363 & 137 \\ 87 & 413 \end{bmatrix}$	$\begin{bmatrix} 393 & 109 \\ 115 & 383 \end{bmatrix}$

Tabela 3. Matriz de confusão

Conclusão

Observando os resultados obtidos, pode-se perceber que de fato a grande dimensionalidade inicial da base de dados era um obstáculo para o classificador. O sentimento de uma sentença varia muito dependendo do contexto em que ela

é inserida, tornando menos previsível a classificação para maiores quantidades de atributos. A redução de dimensionalidade da matriz de características conseguiu um aumento na taxa de acerto em ambas as etapas (treino e teste), fornecendo características mais densas e em menor quantidade. Um ponto a ser notado também é o fato de que pessoas diferentes exprimem as suas opiniões de maneiras diferentes, o que atrapalha a previsibilidade do classificador.

Referências

Allen, J. (1995). *Natural Language Understanding (2Nd Ed.)*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.

Colen, W. (2018). Sistemas de processamento de linguagem natural na prática.

Joshi, A. K. (1991). Natural language processing. *Science*, 253(5025):1242–1249.

Soares, M. V. B., Prati, R. C., and Monard, M. C. (2009). Improvement on the porter stemming algorithm for portuguese. *IEEE Latin America Transactions*, 7(4):472–477.

Tan, S. (2006). An effective refinement strategy for knn text classifier. 30:290–298.