Análise de sentimentos sobre reviews do IMBD

Mateus Rangel
Instituto de Computação
Universidade Federal Fluminense
Rio de Janeiro, Brazil
mateusrangel@id.uff.br

Rafael Mynssem
Instituto de Computação
Universidade Federal Fluminense
Rio de Janeiro, Brazil
rafaelmynssem@id.uff.br

Leonardo Thimoteo
Instituto de Computação
Universidade Federal Fluminense
Rio de Janeiro, Brazil
leonardo_thimoteo@id.uff.br

RESUMO

Este artigo apresenta o desempenho de algoritmos de aprendizado de máquina supervisionados frente a mesma tarefa: classificação de reviews a partir do processamento de linguagem natural. O pré-processamento da base de dados e o treinamento de algoritmos supervisionados de classificação compõe o processo de classificação da polaridade dos reviews avaliados. Ainda compara métricas como Accuracy, Recall, Precision e F1-score obtidas pelos algoritmos performados.

Palavras-Chave

Machine learning, PLN, Sentiment analysis, Text classification.

1. INTRODUÇÃO

Hoje em dia produzimos um volume de dados de uma ordem de grandeza quase inimaginável, cerca de 2.5 quintilhões de bytes por dia. Um quintilhão equivale a 10^{18}[1]. Esses dados carregam muitas vezes informações relevantes que podem ser utilizadas para benefícios de negócios e indivíduos, por exemplo.

As redes sociais são grandes responsáveis por esses números, pela intensidade de interações propostas por essas aplicações. Imaginemos uma situação hipotética na qual um influenciador digital do Instagram com milhões de seguidores deseja entender como as pessoas reagem às suas publicações para além da métrica quantitativa, os likes. Seria necessário avaliar os comentários de uma foto, por exemplo. Com milhões de seguidores os comentários de uma publicação podem chegar a ordem de dezenas e até centenas de milhares, portanto, analisá-los um a um torna-se uma tarefa humanamente inviável.

Conseguir trazer para a dimensão quantitativa um dado qualitativo, como um comentário, capturando quantos foram positivos e negativos, por exemplo, seria um ganho significativo de informação.

Para atacar esse problema, surgem estudos sobre análise de sentimento.

ACM ISBN

De acordo com (Liu, 2012), a análise de sentimentos é a área de estudo que analisa as opiniões, sentimentos, avaliações, apreciações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e todos os seus atributos relacionados.

O objetivo dessa técnica é obter, de forma automática, a polaridade de um texto ou sentença.

- Muita informação (relevante e irrelevante)
- Fontes e formatos diversos
- Dados não estruturados
- Variável temporal
- Dados que não seguem a norma culta da escrita (Gírias, abreviaturas, oralidade, marcas textuais típicas da web, etc.)
- Fatos vs. boatos
- Malícia, manipulação

Neste artigo propomos uma análise de desempenho de algoritmos de aprendizado de máquina supervisionados sobre uma mesma tarefa: classificação de reviews de filmes em positivo ou negativo. Para o trabalho utilizou-se uma base de dados construída a partir de dados do Internet Movie Database (IMDB), grande autoridade na rede quando o assunto é filmes, séries e conteúdos análogos. A base contem entre outros atributos, os reviews de filmes, que são o objeto de estudo. Para cumprir o processamento da linguagem natural houve um passo anterior que foi o pré-processamento da base de dados para a então performance dos algoritmos. A partir disso pudemos então analisar os resultados obtidos a partir de métricas familiares a esse contexto como Accuracy, Recall, Precision e F1-score.

Este trabalho está organizado da seguinte forma: na Seção 2 são apresentadas as características do DataSet utilizado, que chamaremos de Corpus. A seção 3 trará um resumo sobre a tarefa de pré-processamento do Corpus. Na seção 4 são apresentados os algoritmos de machine Learning supervisionados utilizados nos experimentos. Na seção 5 temos o desempenho e os resultados obtidos nos experimentos. Na seção 6, e última, apresenta-se o conclusão e os trabalhos futuros.

2. CORPUS

Por estarmos trabalhando no contexto de processamento de linguagem natural chamaremos a nossa base de dados de Corpus.

Essa base foi obtida a partir do Internet Movie Database (IMDB), fonte mais popular e autoritária do mundo para conteúdo de filmes, programas de TV e celebridades. No IMDB há milhares de ratings e reviews feitos por usuários de diversos lugares do mundo. O idioma majoritário é o inglês e o corpus para esta proposta apresenta somente informações nesse idioma.

Para os experimentos foram utilizados dois corpus, um de treinamento e outro de teste. Ambos com as mesmas características que serão descritas abaixo:

- Origem: IMDB
- Idioma: inglês
- Número de instâncias: 25.000
- Classes: positivo(1) e negativo(0)
- Número de classes: 2
- Classes balanceadas: 12.500 positivas e 12.500 negativas

Há três campos de dados nos arquivos presentes no corpus: id (um identificador para cada review), sentiment (1 para reviews positivos e 0 para reviews negativos) e review (o texto de cada review).

	id	review	sentiment
0	5814_8	With all this stuff going down at the moment w...	1
1	2381_9	\The Classic War of the Worlds\" by Timothy Hi...	1
2	7759_3	The film starts with a manager (Nicholas Bell)...	0
3	3630_4	It must be assumed that those who praised this...	0
4	9495_8	Superbly trashy and wondrously unpretentious 8...	1

Figura 1: Amostra da base de dados antes do pré-processamento.

É importante ressaltar, apesar de abstrairmos o processo de construção do corpus, que por se tratar de um crawling, as informações extraídas possuem TAGs HTML, por exemplo. Dadas essas características do Corpus, vale frisar, ainda, que nessa construção os reviews não estão associados a um filme. A análise está restrita portanto ao processamento dos reviews somente.

3. PRÉ-PROCESSAMENTO

O pré-processamento de dados é uma tarefa muito importante no processo de aprendizado de máquina, pois faz com que a base de dados fique com menos dados irrelevantes ou redundantes fazendo com que tais ruidos sejam mitigados.

No nosso trabalho, primeiro retiramos os reviews cujo conteúdo era nulo, em seguida fizemos com que todas as palavras fossem convertidas para minusculo, retiramos tags HTML, fizemos a tokenização das frases, removemos stopwords e fizemos a lemmatizacao.

As tarefas de tokenização, remoção de stop words e lematização serão detalhadas a seguir.

3.1 Tokenization

Dado uma sequência de caracteres, tokenização é o processo que os separa em partes, chamada tokens. No idioma inglês, a tokenização separa as palavras por espaços e ignora pontuações que não sejam pontos.

3.2 Remoção de stop words

Stop words se referem as palavras mais comuns em um idioma, por serem exaustivamente escritas em um texto, elas normalmente são retiradas do corpus antes de irem para a etapa de modelagem por não adicionam nada à tarefa de análise de sentimentos. Exemplos de stop words no idioma inglês são: a, an, the, and, but, if as, at. Por exemplo, a frase "i like reading, so I read" viraria "like, reading, read"

3.3 Lematização

Lematização é uma técnica para encontrar a raiz da palavra, ignorando os tempos verbais. Essa técnica considera o contexto em que a palavra se encontra e a converte para sua forma base. Por exemplo, a frase "eu encontrei" viraria "eu encontrar"

O resultado do pré-processamento sobre o corpus pode ser avaliado seguir.

	id	review_final	sentiment
0	5814_8	['stuff', 'go', 'moment', 'm]', 'start', 'list...	1
1	2381_9	['classic', 'war', 'timothy', 'hines', 'entert...	1
2	7759_3	['film', 'start', 'manager', 'nicholas', 'bell...	0
3	3630_4	['must', 'assume', 'praise', 'film', 'great', ...	0
4	9495_8	['superbly', 'trashy', 'wondrously', 'unpreten...	1

Figura 2: Amostra da base de dados depois do pré-processamento.

4. ALGORITMOS UTILIZADOS

Para o treinamento dos algoritmos utilizados, dividimos nossa base de dados em: 70% para o treino e 30% para o teste.

A fim de podermos classificar as avaliações, utilizamos três modelos: Naive Bayes, *Random Forest* e máquina de vetores suporte (SVM). Faremos uma breve descrição destes algorítmos assim como mostraremos as suas respectivas matrizes de confusão obtidas depois de implementadas.

4.1 Naive Bayes

Naive Bayes é um algorítmo classificador baseado no Teorema de Bayes. Em aprendizado de máquina, o algoritmo é muito utilizado para análise de sentimentos.

Como características positivas, podemos citar a relativa velocidade do algositmo. Além disso, o mesmo necessita de uma quantidade pequena de dados para obter boa precisão na classificação. Como pontos negativos, o algoritmo não considera as correlações entre as *features* dos dados.

4.2 Random Forest

O *Random Forest* cria várias árvores de decisão e as combina para obter uma predição com maior *accuracy* e mais estável. Como exemplo simples da criação de uma Random Forest baseado em duas árvores de decisão, temos a imagem 4. Uma das grandes vantagens de sua utilização está no fato

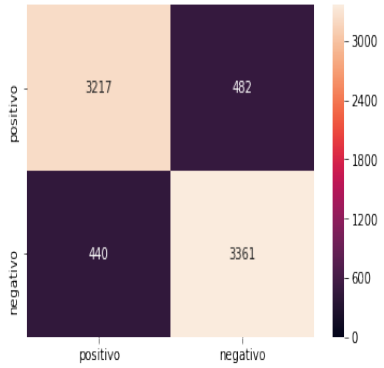

Figura 3: Matriz de confusão do modelo Naive Bayes utilizado.

de que o algoritmo pode ser tanto utilizado para classificação quanto para regressão. Também podemos citar sua fácil interpretação (visto que é baseada em árvores de decisão). Como um ponto negativo, para termos uma boa precisão precisamos de um número elevado de árvores de decisão e isso pode deixar o algoritmo lento para predições em tempo real.

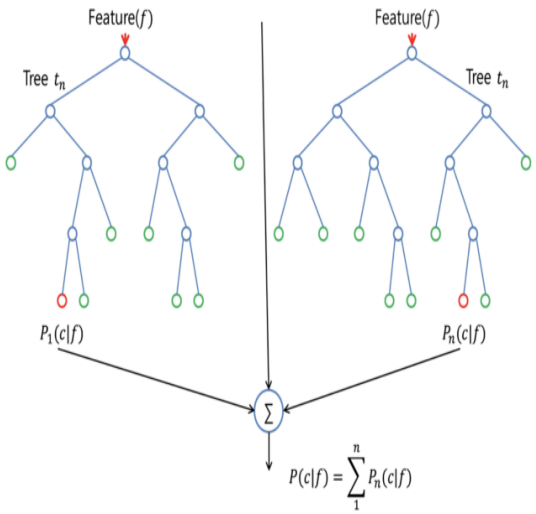

Figura 4: Exemplo de Random Forest criada a partir de 2 árvores de decisão.

4.3 Máquina de vetores suporte (SVM)

A ideia principal do SVM (*support vector machine*) é obter o hiperplano a fim de poder superar, da melhor forma possível, os dados de duas classes distintas.

Entretanto, para obter o melhor hiperplano, antes é necessário definir bem a margem de cada classe. O hiperplanoé, então, obtido através da maximização da distância entre as margens das classes.

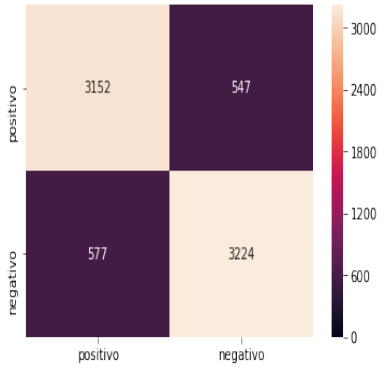

Figura 5: Matriz de confusão do modelo *Random Forest* utilizado.

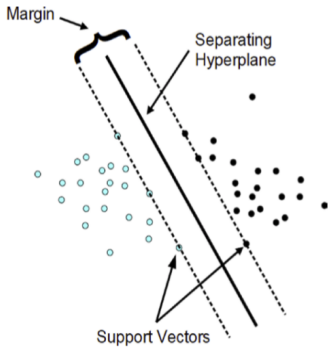

Figura 6: Exemplo das margens separando duas classes. Na imagem, podemos também observar o hiperplano encontrado.

Como pontos positivos do algoritmo, podemos citar o fato do mesmo conseguir lidar bem com os *outliers*. Em alguns momentos, o algoritmo pode até mesmo desconsiderá-los. Se existe uma região, no espaço de representação, que separa as classes, então, é possível obter o hiperplano. Entretanto, o SVM é computacionalemnte complexo podendo, até mesmo, ter ordem cúbica.

5. RESULTADOS

Nessa seção, baseado nos resultados obtidos pelos algoritmos performados, vamos avaliar o desempenho de cada um deles a partir das seguintes métricas de desempenho: *Accuracy*, *Precision*, *Recall*, F1-Score.

A *Accuracy* indica a performance geral do modelo, dentre todas as classificações, quantas foram feitas corretamente.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Já a métrica *precision* é capaz de avaliar, das instâncias que foram classificadas como supostamente da classe correta, quantas realmente estavam corretas, e por isso:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Essa métrica considera a frequência que o classificador en-

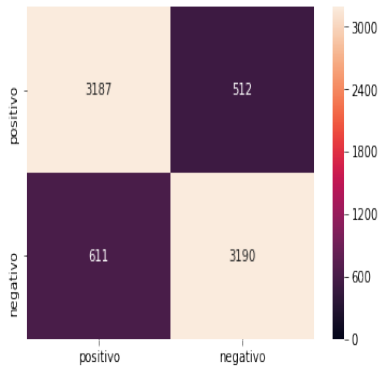

Figura 7: Matriz de confusão do modelo SVM utilizado.

contra os exemplos de uma classe corretamente. Isso significa que, quando a instância realmente pertence à Classe X, o quão frequente é feita a predição da classe X de forma correta.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Essa métrica combina *precision* e *recall* de modo a trazer um número único que indique a qualidade geral do seu modelo. Trata-se de uma média harmônica.

$$F1 - score = 2 * \frac{Precision * Recall}{(Precision + Recall)} \tag{4}$$

Legenda: Verdadeiro Positivo = (TP) Verdadeiro Negativo = (TN) Falso Positivo = (FP) Falso Negativo = (FN)

A *recall* e a *accuracy* são métricas úteis capazes de revelar problemas no modelo que podem ter sido mascarados pela generalização da Acurácia. Um exemplo seria o desbalanceamente de instâncias entre as classes. O textitF1-score foi um recurso selecionado por permitir que através de apenas uma métrica seja possível avaliar o desempenho de duas outras, da textitrecall e da *precision*, confirmando a eficiência do modelo.

	Accuracy	F1 Score	Precision	Recall
Naive Bayes	85.013	85.012	85.016	85.011
Random Forest	85.027	85.027	85.042	85.041
SVM	87.707	87.702	87.697	87.713

Tabela 1: Tabela com os resultados dos scores obtidos para os três algoritmos utilizados.

A partir dos resultados resumidos na tabela 1, concluímos que os resultados, em destaque, de todas as métricas analisadas foram mais satisfatórios no algoritmo SVM (Support Vector Machine).

6. CONCLUSÃO E TRABALHOS FUTUROS

Avaliamos neste trabalho a performance de diferentes algoritmos de aprendizado de máquina supervisionado sobre a mesma tarefa: análise de sentimento. Exemplificamos passos importantes do processo como o pré-processamento da base de dados para extrair resultados mais relevantes. Concluímos, através das métricas avaliadas, que os resultados foram bastante satisfatórios se considerarmos que a classe majoritária corresponde a 50% do corpus e os resultados de acurácia, de todos os algoritmos, superaram significativamente esse valor, chegando a 87%.

Como trabalho futuro pretende-se avaliar o desempenho de algoritmos de aprendizado não supervisionado e ainda utilizar técnicas mais sofisticadas de pré-processamento da base de dados, como Bag of words.

7. REFERÊNCIAS

Bing, Liu. "Sentiment Analysis and Opinion Mining", Morgan Claypool Publishers, 2012

Marr, Bernard. "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read." Forbes, Forbes Magazine, 11 Mar. 2019,

www.forbes.com/sites/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/4db68fb460ba.

Borba, Vitor. Métricas De Avaliação: Acurácia, Precisão, Recall... Quais As Diferenças?

https://medium.com/@vitorborbarodrigues/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c

Tyagi, P., Tripathi, R. C. (2019). A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data. SSRN Electronic Journal. doi:10.2139/ssrn.3349569

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze .(2009). An Introduction to Information Retrieval. Cambridge University Press.