Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Lucas Vinicio Maldonado

Classificação de tweets relevantes em situação de desastre

Curitiba
2022

Lucas Vinicio Maldonado

Classificação de tweets relevantes em situação de desastre

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Luiz Eduardo S. Oliveira

Curitiba
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Classificação de tweets relevantes em situação de desastre

Lucas Vinicio Maldonado[1]
Luiz Eduardo S. Oliveira[2]

Resumo

Entre as diversas categorias de tarefas atualmente viáveis para o Processamento de Linguagem Natural, a classificação de texto é uma das de maior destaque pela sua versatilidade entre domínios e aplicações especificas. Na consolidação global da utilização das redes sociais, nas quais inúmeras pessoas compartilham informações das mais diversas naturezas, encontra-se a aplicação de distinguir quais fatos são reais. No meio desse escopo, a classificação efetiva de noticias relativas a desastres e incidentes possui particular relevância, de forma a governos e órgãos de auxilio atuarem com a maior agilidade possível com o objetivo de reduzir e amenizar danos. Neste artigo são comparados o desempenho de 3 algoritmos clássicos de machine learning e um de deep learning (BERT), que possui representações contextuais treinadas em redes neurais profundas, na classificação de tweets relativos a desastres reais. Os resultados apresentados por meio de métricas de desempenho demonstram que o BERT supera significativamente os demais algoritmos nesta tarefa.
Palavras-chave: Processamento de Linguagem Natural, Classificação de Texto, Detecção de desastres, BERT, redes sociais.

Abstract

Among the high diversity of categories related to practicable tasks nowadays in Natural Language Processing, text classification is one of the highlights by the versatility amid domains and specific applications. By the global usage consolidation of social networks, where innumerable people share information about all sorts of nature, we find the application of determining which facts are real. In this scope, the effective classification of news related to disasters and incidents has a particular relevance, in order to governments and rescue agencies act as soon as possible to reduce damages. In this article the performance of 3 classical machine learning algorithms and one deep learning (BERT), with contextual representations trained on deep neural networks, are compared for the classification of tweets related to real disasters. The results by performance metrics shows that BERT overcomes significantly the remaining algorithms in this task.
***Keywords**: Natural Language Processing, Text Classification, Disaster Detection, BERT, Social Networks.*

1 Introdução

A classificação de texto é uma das mais populares tarefas do Processamento de Linguagem Natural, sendo atualmente utilizada nas mais diversas industrias como saúde, mídias sociais, direito e marketing, entre outras. Esta tarefa por si é um caso especifico dos modelos de classificação em geral de Machine Learning, os quais procuram categorizar dados em uma ou mais classes[1]. Dentre as inúmeras aplicações especificas para textos, encontram-se classificadores de spam para e-mails, analise de sentimentos e detectores de noticias falsas(fake news).

A detecção de informações falsas tem dentro de suas diversas possibilidades a identificação de comunicações sobre desastres naturais e outra emergências. Pela atual presença massiva de smartphones e uso das redes sociais, estas informações podem ser compartilhadas praticamente em tempo real pela observação direta das pessoas. Dentre as diversas redes mais utilizadas, o Twitter se tornou um dos principais meios de comunicação para noticias e, por consequência, de dados a serem possivelmente analisados e classificados. Governos, jornais e organizações de atendimento a desastres são alguns dos grupos que estão constantemente monitorando esta rede para pronta atuação e divulgação. [2]

O objetivo deste trabalho é comparar a efetividade da classificação destes *tweets*(mensagens com até 280 caracteres publicados no Twitter) por meio de 4 algoritmos, entre eles o modelo BERT que é uma das aplicações de Transfer Learning - a técnica atualmente considerada o estado da arte para PLN.

Na seção Conjunto de Dados deste artigo é descrita a base de dados com sua origem, características das variáveis e respectiva analise exploratória. Em Metodologia são descritos os algoritmos utilizados, o pré-processamento dos textos, a extração de características e os parâmetros utilizados. Em Resultados experimentais são apresentadas as métricas utilizadas, os resultados obtidos e as devidas comparações. Finalmente, em Conclusões são apresentados comentários e observações que

[1]Aluno do programa de Especialização em Data Science & Big Data, lucasvinicio@outlook.com.
[3]Professor do Departamento de Informática - DINF/UFPR. luiz.oliveira@ufpr.br.

sumarizam os experimentos.

2 Conjunto de Dados

O conjunto de dados utilizado neste estudo consiste em 11370 tweets em inglês classificados de forma binária como relevante ou não sobre desastres de diversas naturezas, como incêndios florestais, a pandemia do coronavirus, ações militares e acidentes urbanos.

O dataset público foi disponibilizado pela Kaggle, uma plataforma amplamente conhecida para estudantes e pesquisadores em Machine Learning. A primeira versão divulgada em 2020 faz parte da competição Natural Language Processing with Disaster Tweets, tendo recebido atualizações para o incremento da base. Esta primeira versão foi criada e classificada pela Appen. Além da classificação binária manual, o conjunto possui para cada tweet uma palavra chave para o tipo de incidente e a localização na maioria dos registros.

2.1 Analise descritiva

De forma a conhecer mais o conjunto antes de começar as etapas de pré-processamento, é realizada uma analise exploratória que pode detalhar mais suas características. 81,4% dos tweets são classificados como não sendo desastres e 18,6% relativos a reais incidentes. É um dataset claramente desbalanceado, o que tende a gerar algoritmos que prevem com maior assertividade tweets não reais. A Tabela 1 apresenta alguns exemplos com sua respectiva classificação.

Tabela 1: Exemplos de tweets no conjunto.

Tweet	Target
Puerto Rico hit by another 5.9 magnitude aftershock - KYMA https://t.co/wW3CdsvDV2 (Porto Rico atingido por outro abalo secundário de 5.9 de magnitude)	1 (Incidente real)
Australian Open matches to be confined to indoor courts if conditions turn hazardous | Article [AMP] | Reuters https://t.co/0jQtRXoFen (As partidas do Australian Open serão limitadas a quadras cobertas se as condições se tornarem perigosas)	0 (Não um incidente)

Os tweets tem em média 107 caracteres e 17 palavras. A Figura 1 mostra as 12 palavras com maior frequência dentre os tweets positivos, após pré-processamento que incluiu remoção de stopwords e remoção de tags html.

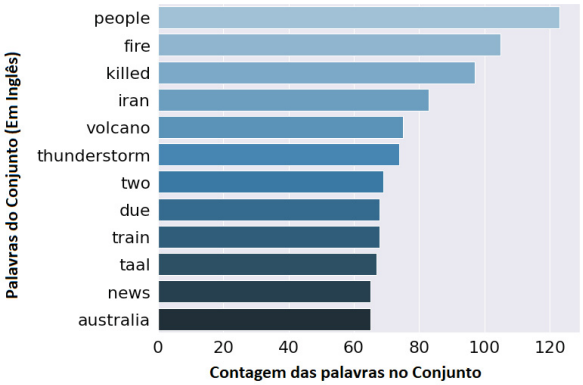

Figura 1: Palavras em inglês mais frequentes dentre os tweets relativos a desastres.

3 Metodologia

Foram selecionados 3 algoritmos clássicos de Machine Learning e 1 de Deep Learning baseado em transformers, o BERT (Bidirectional Encoder Representations from Transformers).

3.1 Naïve Bayes

Segundo [3], Naïve Bayes é um classificador probabilístico que utiliza o teorema de Bayes para estimar a probabilidade condicional de cada feature relativo aos tokens de um determinado texto para cada categoria, dada a ocorrência dessa feature na categoria. São então multiplicadas todas as probabilidades de todas as features e determinada a categoria com base na maior probabilidade.

3.2 Regressão logística

Uma das principais bases para modelos supervisionados de aprendizado de máquina, a regressão logística tem por principal diferença com Naïve Bayes o fato de ser um classificador discriminativo, ao contrário de generativo. Desta forma, a regressão logística procura atribuir pesos que auxiliem de forma mais eficiente a separar as classes analisadas, ainda que talvez não seja possível gerar um exemplo destas. O modelo utiliza por base a função sigmoidal(ou logística) para construir um vetor com pesos que ajudem a diferenciar entre duas classes as palavras que compõem um documento. [4]

3.3 Support Vector Machine (SVM)

Uma máquina de vetores de suporte, igualmente um modelo discriminativo, é desenvolvida para encontrar fronteiras em planos em planos hiperdimensionais, levando a esta discriminação ser a máxima possível. Ao contrário da regressão logística, SVM's são capazes de gerar separações não lineares. Sua maior vantagem é a adaptatividade a variações e erros que os dados podem possuir. Todavia, este modelo leva muito mais tempo em

ser treinado, não sendo escalável para grandes volumes de dados. [3]

3.4 BERT

Os algoritmos citados previamente utilizam representações de características que determinam um conjunto fixo para cada palavra. De forma a captar o contexto destas, em 2018 foi proposto pelos autores de [5] um modelo que utiliza representações contextuais das palavras, o Bidirectional Encoder Representations from Transformers(BERT), do inglês Representações de codificador bidirecional de Transformers. Este modelo, e suas diversas variantes por consequência, foi desenvolvido de forma a ser pré-treinado em representações bidirecionais de textos sem classificação, para depois ser aperfeiçoado por meio das últimas camadas com textos classificados para tarefas específicas [1].

Conforme os autores de [1], assim como por meio de diversas outras publicações, BERT costuma superar algoritmos clássicos que utilizam representações não-contextuais(como Bag of words e TF-IDF). Todavia, é relevante analisarmos se para esta base significativamente desbalanceada, com uma origem que costuma ter alto uso de abreviações e gírias e para o contexto específicos de identificar eventos relacionados a desastres, o modelo vai superar os outros citados.

3.5 Pré-processamento e extração de características

De forma geral, existem etapas em comum que devem ser implementadas antes de criar e utilizar modelos de processamento de linguagem natural. Para todos os modelos utilizados neste estudo, foram aplicadas com a biblioteca BeautifulSoup e de expressões regulares a remoção de tags html, links, caracteres correspondentes a emojis e outras expressões, além de normalizar todas as palavras para minusculas. Apenas para os modelos clássicos foi também aplicado a remoção de stopwords (ou palavras vazias) - formas como artigos e preposições que não incrementam no significado do contexto de cada documento.

Para os modelos Naive Bayes, Regressão logística e SVM foram utilizadas as representações de características Bag of Words (BOW) e TF-IDF(do inglês term frequency–inverse document frequency). Esta primeira consiste em contabilizar a ocorrência de cada palavra de um documento, dentro do conjunto total de palavras, ignorando a frequência, ordem e contexto. A distancia euclidiana entre cada vector pode demonstrar a similaridade semântica entre os documentos. Entretanto, ainda que seja uma representação de fácil implementação, esta não captura qualquer contexto dos documentos.

Por outro lado, a representação com TF-IDF procura dar maior peso a palavras que sejam particularmente mais importantes para determinado conjunto. Isto ocorre com base na multiplicação de duas medidas: TF (Frequência do termo), que estabelece maior relevância para as palavras que mais aparecem em cada documento, e IDF (inverso da frequência nos documentos), que mede a importância de um termo entre todos os documentos de um conjunto.

Por último, para implementação e ajuste para esta classificação do modelo BERT, foi escolhido a instância "bert-base-uncase", que ja possui os embeddings treinados com 12 camadas de transformers e 768 camadas ocultas. Os embedding criados para este modelo seguem o método WordPiece tokenization.

3.6 Métricas

No estudo foram empregadas duas métricas para avaliar os modelos: F1-score, derivada da sensibilidade e exatidão, e Área sob a curva ROC (AUC). Não foi optado por incluir a acurácia devido ao conjunto de dados ser desbalanceado. Para este experimento é considerado como verdadeiro positivo (TP) os tweets que são realmente desastres e foram previstos como tais. Já os falsos positivos (FP) são tweets que não eram determinados como desastres, porém o modelos categorizou como reais. Verdadeiro-negativo (TN) e falso-negativo (FN) seguem a mesma lógica.

Com estes conceitos definidos, temos que a *sensibilidade* é a proporção de TP sobre o total de resultados marcados como positivos, medindo dessa forma a capacidade de detectar com sucesso resultados positivos. Por outro lado, a *exatidão* ou *precisão* avalia a quantidade de verdadeiros positivos sobre o total de positivos.

$$\text{Sensibilidade: } R = \frac{TP}{TP + FN}.$$

$$\text{Exatidão: } P = \frac{TP}{TP + FP}.$$

Ao combinar sensibilidade e precisão, obtemos a *F1* pela média harmônica entre elas. Dessa forma, F1 é formada pelo mesmo peso para ambas as métricas. Caso os valores obtidos para F1 sejam altos, significa que tanto sensibilidade quanto exatidão tiveram resultados altos - sendo análogo para resultados baixos. Todavia, se obtivermos um resultado médio para F1, apenas uma das métricas que o compõe teve um resultado alto.

$$\text{F1-score: } F1 = 2 \cdot \frac{P \cdot R}{P + R}.$$

A Curva Característica de Operação do Receptor(do inglês Receiver Operating Characteristic Curve - ROC) é um gráfico que mostra o desempenho de classificação de um modelo a partir da variação de seus limites. É construído plotando a sensibilidade, também conhecida como taxa de verdadeiros positivos, contra a taxa de falsos negativos. Esta é definida como $FPR = 1 - Sensibilidade$.

A Área sob a curva ROC(AUC), que mede de forma integral a área de (0,0) até (1,1), demonstra que tão bem o modelo consegue distinguir entre ambas as classes analisadas. [6]

3.7 Ajuste dos modelos

O principal parâmetro modificado para os modelos clássicos foi o número de features máxima para cada um. Foram encontradas como próximas ao ideal 6000 para Naive Bayes e Regressão Logística e, por outro lado, 8000 para SVM. Dentre os hiperparametros testados no modelo BERT, a melhor performance foi com batch size igual a 16, taxa de aprendizado de 2e-5 e realizado com 3 épocas de treinamento.

De forma a obter um melhor desempenho para ambas as classes foi aplicado na separação de base de treino e teste a amostragem estratificada.

4 Resultados experimentais

Os resultados de F1-Score e AUC são apresentados na Tabela 2 para todos os modelos de machine learning utilizados e para todas as representações de características. Dentre os modelos clássicos, a Regressão logística foi a que obteve melhor desempenho para a representação BOW. A utilização de TF-IDF não apresentou significativas melhorias, o que pode ser devido a alta variância de incidentes no conjuntos. Por consequência, não há um dicionario extenso de palavras que ajudem a identificar e diferenciar os documentos reais.

A utilização do modelo BERT com sua representação contextual demonstrou melhorias expressivas para estas métricas. A constituição destes embeddings com um pré-treinamento extenso em outros conjuntos mostra ser muito vantajoso.

Tabela 2: F1-Score e AUC para os modelos utilizados

Modelo	F1	AUC
Naïve Bayes (BOW)	0,64	0,86
Regressão Logística (BOW)	0,66	0,89
SVM (BOW)	0,60	0,86
Naïve Bayes (TF-IDF)	0,46	0,87
Regressão Logística (TF-IDF)	0,63	0,87
SVM (TF-IDF)	0,65	0,88
BERT	0,71	0,93

São apresentados a seguir - nas Figuras 2, 3 e 4 - as matrizes de confusão para os os 3 modelos com melhor desempenho geral para cada representação de características. Ainda que para a representação com TF-IDF as medidas apresentadas previamente sejam superiores para SVM, a matriz de confusão indicou que a efetividade para os valores reais é superior para a Regressão Logística, de 0,68 a 0,70. Neste caso, sendo os demais valores próximos, este último modelo é considerado com o melhor desempenho para TF-IDF.

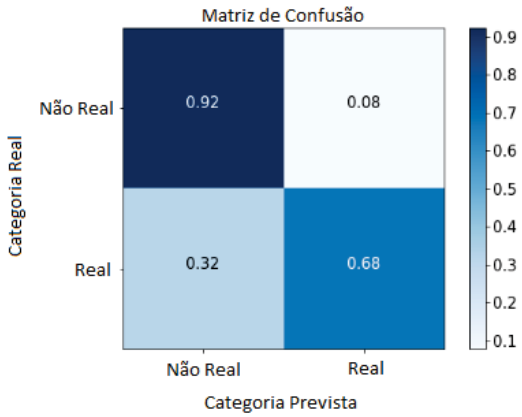

Figura 2: Matriz de confusão para o modelo Regressão Logística com BOW.

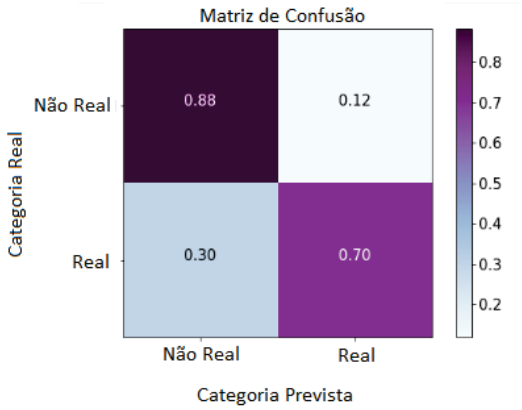

Figura 3: Matriz de confusão para o modelo Regressão Logística com TF-IDF.

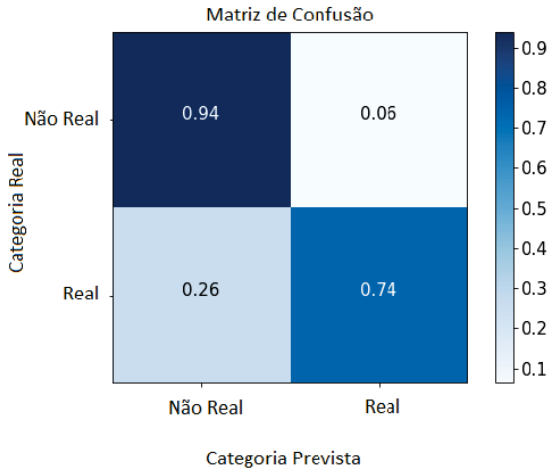

Figura 4: Matriz de confusão para o modelo BERT.

5 Conclusões

Neste artigo foram descritos a utilização e desempenho na classificação de tweets relativos a desastres por meio

de diferentes modelos de machine learning e representações de características. Os resultados experimentais demonstram que a utilização do modelo de redes neurais profundas com representação contextual supera os demais modelos.

É importante notar que os modelos clássicos apresentaram resultados satisfatórios que podem ser aprimorados com outras representações de características e parâmetros. Estas possibilidades são particularmente relevantes visto que, ainda que os modelos de deep learning no estado da arte superam na efetividade geral, a produtização destes modelos mais simples é amplamente mais realizável. BERT (e suas diversas variantes), assim como outros modelos mais avançados, requer um alto poder de processamento computacional para treinamento completo, limitado a nível empresarial e instituições de ensino com recursos disponíveis.

Trabalhos futuros podem investigar as possíveis melhorias para modelos clássicos, combinar diversos modelos, assim como explorar outros modelos mais recentes como GPT-3[7], com aproximadamente mais de 100 vezes os parâmetros do modelo BERT utilizado.

Referências

[1] González-Carvajal e Garrido-Merchán. *Comparing BERT against traditional machine learning text classification*. 2021.

[2] Ashktorab Zahra et al. *Mining Twitter to Inform Disaster Response*. ISCRAM, 2014.

[3] Sowmya Vajjala et al. *Practical Natural Language Processing*. O'Reilly Media, Inc., 2020.

[4] Dan Jurafsky e James H. Martin. *Speech and Language Processing*. 2021.

[5] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018.

[6] Google Machine Learning Crash Course. *Classification: ROC Curve and AUC*. 2021.

[7] Brown et al. *Language Models are Few-Shot Learners*. 2020.