ANÁLISE DE SENTIMENTOS EM COMENTÁRIOS DE CONSUMIDORES: UMA ABORDAGEM COM EMBEDDINGS E MODELOS DE APRENDIZADO DE MÁQUINA

Marilia Thomaz Fernandes da Silva[1], Fabiano Oliveira Costa Prado[2], Sandra Cristina Costa Prado[3]

[1] Discente do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas / marilia.silva9@fatec.sp.gov.br
[2] Docente da Universidade Estadual Paulista / fabiano.prado@unesp.br
[3] Docente do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas / sandra.prado01@fatec.sp.gov.br

RESUMO
Neste TCC foi realizada uma análise de sentimento em comentários de consumidores de produtos para bebês, utilizando uma representação vetorial chamada de embeddings. Foram utilizados dois modelos de aprendizado de máquina para realizar a análise de sentimento: um modelo linear (SGDClassifier) e um modelo não-linear (Random Forest). O modelo linear apresentou uma perfomance preditiva ligeiramente menor que o modelo não-linear. Além disso, os resultados obtidos foram comparados com um Trabalho de Conclusão de Curso apresentado no VIII CONGRESSO DE TRABALHOS DE GRADUAÇÃO da Faculdade de Tecnologia de Mococa Vol.8 N.2 A.2021, onde a análise de sentimento foi realizada utilizando a representação de saco de palavras (bag of words). Os resultados indicam que a representação dos comentários por meio de embeddings apresentou uma melhoria significativa na acurácia da classificação de sentimentos em comparação com a representação por meio de saco de palavras e, também, foi possível identificar e propor uma utilização hibrida dos dois modelos, garantindo simultaneamente otimização nas inferências e nível de predição alto.
Palavras-chave: Aprendizado de máquina; Processamento de Linguagem Natural; Embeddings

1 INTRODUÇÃO

A análise de sentimentos (BIRD, 2009) é uma área do processamento de linguagem natural (PLN) que visa identificar e extrair informações subjetivas dos textos, determinando se o sentimento expresso é positivo, negativo ou neutro. Esta técnica tem ampla aplicação em diversos domínios, incluindo a análise de produtos, monitoramento de redes sociais, suporte ao cliente e pesquisa de mercado. No presente trabalho, a análise de sentimentos foi feita em comentários sobre produtos de bebê disponibilizados pela empresa Amazon (AMAZONBABY, 2024).

Neste projeto, foi adotada uma abordagem que utiliza representações avançadas de texto, conhecidas como embeddings, geradas pelo modelo Sentence

Transformers (REIMERS, 2019). Essas representações vetoriais capturam de forma eficaz as nuances semânticas e contextuais dos textos, superando a tradicional técnica de saco de palavras (bag of words) em termos de riqueza de informação e desempenho preditivo.

Foram utilizados dois modelos de aprendizado de máquina para a análise de sentimentos: um modelo linear (BISHOP, 2006) e o modelo Random Forest (BREIMAN, 2001). O modelo linear, pela sua simplicidade e eficiência, oferece uma solução prática e rápida para a classificação de sentimentos. Por outro lado, o modelo Random Forest, com sua capacidade de modelar relações mais complexas, apresenta uma abordagem robusta e flexível, capaz de lidar com a variabilidade dos dados de forma mais eficaz.

Um aspecto relevante deste trabalho é a comparação com estudos anteriores que utilizaram diferentes técnicas de representação textual, como o bag of words (BIRD, 2009). A comparação visa demonstrar as vantagens e melhorias proporcionadas pelo uso de embeddings na análise de sentimentos.

Além disso, este trabalho explora a combinação de diferentes modelos para otimizar a precisão das predições e a eficiência computacional. A abordagem híbrida proposta considera a aplicação seletiva dos modelos de acordo com as características dos dados, maximizando o desempenho geral.

O artigo está organizado da seguinte forma: na seção 2 são apresentados os principais conceitos envolvidos neste trabalho; a seção 3 mostra os passos seguidos na implementação deste projeto; na seção 4 é feita uma breve discussão dos resultados obtidos e comparação com resultados anteriores; a seção 5 é dedicada às conclusões a partir do que foi exposto na seção 4, e, em seguida, são apresentadas as referências que este trabalho utilizou.

2 REFERENCIAL TEÓRICO

Os embeddings de texto são representações vetoriais de palavras, sentenças ou documentos que capturam a semântica e o contexto dos textos em um espaço de alta dimensão. Ao contrário da representação mais elementar como o saco de palavras (bag of words), que tratam as palavras como independentes umas das outras, os embeddings consideram as relações contextuais entre as palavras, permitindo uma representação mais rica e significativa. Neste trabalho, foi considerado como ferramenta para extrair de uma sentença um embedding, o modelo

Sentence Transformers (ou SBERT) que é uma extensão do BERT (Bidirectional Encoder Representations from Transformers) (DELVIN,2019), um dos modelos mais avançados em PLN.

Originalmente o BERT foi projetado para processar e entender palavras em contexto, sendo que o SBERT adapta essa capacidade para trabalhar com frases e sentenças, gerando embeddings de alta qualidade que capturam o significado das sentenças inteiras. Entretanto é necessário que as sentenças tenham um tamanho aproximado de 10 a 40 tokens para que a performance dos embeddings não seja comprometida com uma representação vetorial ruim da sentença.

A principal vantagem dos embeddings é que eles conseguem capturar as similaridades semânticas entre sentenças. Esta capacidade de capturar a semântica contextual é crucial para tarefas de processamento de linguagem natural (PLN) como a análise de sentimentos, onde o significado das palavras pode variar significativamente dependendo do contexto em que são usadas.

Tanto o BERT como o SBERT são modelos de aprendizado de máquina que utilizam redes neurais profundas (deep learning). Especificamente, a conexão entre os neurônios artificias seguem uma arquitetura denominada transformers. No caso do SBERT, ela foi modificada para a tarefa de similaridade entre sentenças de tal forma que combina os embeddings de palavras do modelo BERT e os transforma em embeddings de sentenças através de um pooling específico.

Como todo algoritmo de aprendizado de máquina profundo, o SBERT foi treinado utilizando uma grande quantidade de dados, especificamente foi utilizado um grande conjunto de documentos textuais.

A análise de sentimento, também conhecida como mineração de opinião, é uma tarefa no campo do PLN que visa identificar e extrair informações subjetivas de textos. O objetivo principal é determinar a atitude, emocional ou opinativa, expressa em uma série de palavras, frases ou documentos. Essa tarefa é amplamente utilizada em diversos domínios, como análise de produtos, monitoramento de redes sociais, suporte ao cliente e pesquisa de mercado.

Dentro das 3 principais tarefas (regressão, classificação e clusterização) que um algoritmo de aprendizado de máquina realiza, a análise de sentimento é fundamentalmente uma tarefa de classificação, onde um texto é categorizado em diferentes classes de sentimentos. As classes mais comuns são "positivo", "negativo" e, em alguns casos, "neutro". No contexto deste trabalho, estamos focados em uma classificação binária: "positivo" e "negativo".

Para realizar a análise de sentimento, utilizamos algoritmos de aprendizado de máquina, que são métodos computacionais que aprendem padrões a partir de dados para fazer previsões ou tomar decisões sem serem explicitamente programados para a tarefa específica. A eficiência desses algoritmos depende da qualidade e representatividade dos dados utilizados no treinamento.

Os algoritmos de aprendizado de máquina são ferramentas poderosas que permitem aos computadores aprender e tomar decisões com base em dados. No regime de dados supervisionados, o processo de aprendizado envolve um conjunto de dados de entrada rotulados, onde cada exemplo de entrada é acompanhado de uma saída desejada ou rótulo. Esse regime é chamado de supervisionado porque o processo de treinamento é guiado por exemplos conhecidos, que atuam como um "professor" para o algoritmo, visto que os dados são apresentados na forma de um conjunto de pares constituído de entrada (“pergunta”) e sua respectiva saída (“resposta”).

O primeiro passo no aprendizado supervisionado é a coleta de um conjunto de dados rotulados. Esses dados consistem em pares de entradas e saídas, onde as entradas podem ser qualquer forma de dados, como texto, imagens, ou valores numéricos, e as saídas são os rótulos correspondentes. A preparação dos dados inclui a limpeza, normalização e transformação das entradas em um formato adequado para o algoritmo de aprendizado de máquina. No presente trabalho, os dados de entrada textuais foram transformados em embeddings o que possibilitou a utilização de algoritmos de aprendizado de máquina implementados pelas classes RandomForestClassifier() e SGDClassifier() do pacote sklearn (PEDREGOSA, 2011).

Uma prática comum é dividir o conjunto de dados em três partes: treinamento, validação e teste. O conjunto de treinamento é usado para ajustar os parâmetros do modelo. O conjunto de validação é usado para ajustar hiperparâmetros e prevenir overfitting, que é quando o modelo se ajusta demais aos dados de treinamento e não generaliza bem para novos dados. O conjunto de teste é usado para avaliar a performance final do modelo.

O aprendizado ocorre durante a fase de treinamento, onde o algoritmo de aprendizado de máquina será capaz de prever uma saída a partir de um novo dado de entrada. Este é o processo mais custoso do ponto de vista computacional visto que nesta fase são ajustados os parâmetros internos do algoritmo de forma iterativa com a finalidade de minimizar os erros entre suas previsões e as saídas reais nos dados

de treinamento. Este processo geralmente envolve a otimização de uma função de perda ou custo, que quantifica o erro das previsões do modelo.

Após a fase de treinamento e validação, o modelo é avaliado usando o conjunto de teste. Esta avaliação fornece uma medida imparcial do desempenho do modelo em dados não vistos, permitindo uma estimativa de como o modelo se comportará em situações do mundo real.

3 METODOLOGIA

O projeto foi inteiramente implementado na linguagem Python, utilizando diversas bibliotecas e pacotes específicos para manipulação, transformação e análise de dados textuais. A codificação foi realizada utilizando a IDE Jupyter no formato de notebooks onde foi possível realizar os comentários de uma forma customizada além de executar cada trecho de código separadamente com a finalidade de testar cada parte da implementação de maneira bem controlada. A seguir, a sequência de procedimentos metodológicos adotados é detalhada.

Os dados textuais foram manipulados principalmente com o auxílio do pacote pandas e do módulo tokenize do pacote nltk (BIRD, 2009). Inicialmente foi realizada a eliminação de dados faltantes, onde todos os registros com dados faltantes foram removidos para garantir a integridade do conjunto de dados. Na sequência, os dados passaram por um processo de transformação onde a coluna de avaliação (rating) foi transformada em dados categóricos, sendo categorizada em "positivo" e "negativo". Comentários com avaliações altas foram rotulados como positivos, enquanto avaliações baixas foram rotuladas como negativas. Posteriormente realizamos o que é chamado de balanceamento dos dados que é necessário sempre que há um desequilíbrio entre as classes, visto que havia mais dados da categoria "positivo" do que "negativo". Para tanto, foi aplicado o critério de undersampling na classe majoritária, garantindo um conjunto de dados balanceado, essencial para o treinamento eficaz dos modelos.

O processo de segmentação e tokenização dos dados textuais foi o próximo passo seguido na metodologia deste trabalho, onde os dados textuais foram segmentados em sentenças e, em seguida, foi feita a contagem dos tokens de cada sentença para análise subsequente (BIRD, 2009). A etapa final da transformação dos dados envolveu a vetorização das sentenças utilizando o pacote sentence_transformers, onde cada sentença foi transformada em embeddings,

representações vetoriais de alta dimensão que capturam o significado semântico dos textos. Nos casos em que o texto apresentava mais de uma sentença, os embeddings foram agregados usando pooling segundo o critério de valor médio. Isso resultou em uma única representação vetorial para cada comentário. Cabe ressaltar que foram filtradas sentenças que estavam na faixa de 8 a 40 tokens, pois a representação vetorial neste caso é a melhor possível.

Após a transformação dos dados, o conjunto de dados foi dividido em duas partes, a saber, 20% dos dados tratados foram reservados exclusivamente para testes, visando a avaliação final do modelo. O restante dos dados foi utilizado para treinamento e validação cruzada dos modelos. Os modelos lineares e Random Forest foram implementados através das classes disponíveis no pacote sklearn. Os modelos foram instanciados e treinados utilizando o processo de "grid search" combinado com validação cruzada, ambos oferecidos pelo pacote sklearn. O grid search permite a otimização dos hiperparâmetros dos modelos ao testar diversas combinações possíveis, enquanto a validação cruzada garante uma avaliação robusta do desempenho dos modelos durante o treinamento. Após o treinamento e a validação, os modelos melhor ajustados dos dois algoritmos escolhidos, foram selecionados para a fase de teste. Na fase final, a performance dos modelos selecionados foi avaliada utilizando o conjunto de teste reservado. Tanto o modelo linear quanto o modelo Random Forest foram testados, e suas performances foram comparadas em termos de métricas como acurácia, precisão, recall e F1-score.

A presente metodologia garantiu que os melhores modelos de cada algoritmo (Linear e Random Forest) fossem treinados, validados e testados de maneira satisfatória, resultando em uma análise de sentimento confiável para os comentários sobre produtos de bebê.

Os detalhes da implementação, bem como suas atualizações, podem ser encontrados no seguinte link: https://github.com/MariliaThomaz/Analise-de-Sentimento-.

4 RESULTADOS E DISCUSSÃO

Apesar do modelo linear ser o mais simples dos modelos utilizados neste trabalho, sua performance foi bem satisfatória com um score de mais de 85% de acertos sobre os dados de teste. Além disso, comparativamente, a assertividade do

modelo linear foi melhor que a assertividade do mesmo modelo utilizado em trabalho anterior sobre a mesma base de dados. Isto se deve ao fato das diferentes representações textuais utilizada em cada um dos trabalhos. No presente trabalho, a representação por embeddings através do modelo SBERT se mostra superior à representação utilizada no trabalho anterior que foi a vetorização pela técnica bag of words.

O modelo Random Forest teve uma performance preditiva um pouco melhor, ultrapassando os 90% de acertos com os dados de teste. Esta superioridade apresentada pelo modelo Random Forest se deve ao fato do modelo ser mais flexível que o modelo linear devido a grande quantidade de parâmetros que o algoritmo Random Forest apresenta em comparação com a quantidade de parâmetros do modelo linear. Entretanto, o algoritmo Random Forest é muito mais complexo do ponto de vista das operações computacionais utilizadas para fazer suas predições.

Utilizando recursos de álgebra linear, foi possível perceber que grande parte dos 5% de diferença entre os dois modelos utilizados são associados a dados que se encontram em uma região específica do espaço vetorial dos embeddings, que é a fronteira linear do modelo SGDClassifier. Portanto, em termos de otimização computacional, é possível fazer uso modelo mais simples sempre que o embedding esteja longe da fronteira linear e quando o embedding estiver próximo da fronteira seja utilizado o modelo mais complexo devido sua preditividade maior nesta região.

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma análise de sentimentos sobre comentários de produtos de bebê, utilizando representações textuais avançadas e algoritmos de aprendizado de máquina. A implementação em Python e a utilização de embeddings de texto gerados pelo modelo Sentence Transformers (SBERT) mostraram-se eficazes resultando em modelos preditivos robustos.

Em termos de otimização computacional, é viável adotar uma abordagem híbrida que aproveite as vantagens de ambos os modelos. O modelo linear, sendo mais simples e eficiente, pode ser utilizado para predições quando os embeddings estão longe da fronteira linear. Quando os embeddings estão próximos dessa fronteira, o modelo Random Forest, com sua maior capacidade preditiva, pode ser acionado para fornecer uma classificação mais precisa. Essa abordagem não só melhora a eficiência

computacional, mas também mantém um alto nível de acurácia na análise de sentimentos.

REFERÊNCIAS

AMAZONBABY. https://www.kaggle.com/ronnie3rg/amazon-baby-sentiment-analysis Acesso em: 2 jun. 2024.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. ArXiv preprint arXiv: 1810.04805, 2019. Disponível em: https://arxiv.org/abs/1810.04805. Acesso em: 13 jun. 2024.

REIMERS, Nils; GUREVYCH, Iryna. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. arXiv preprint arXiv:1908.10084*, 2019. Disponível em: https://arxiv.org/abs/1908.10084. Acesso em: 7 jun. 2024.

PEDREGOSA, Fabian et al. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011. Disponível em: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Acesso em: 11 jun. 2024.

BREIMAN, Leo. **Random forests**. Machine Learning, v. 45, n. 1, p. 5-32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: 13 jun. 2024.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. Springer, 2006.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit**. O'Reilly Media, Inc., 2009.