Análise de Sentimento

Um Complemento na Análise de Satisfação dos Usuários

FERRERAS, RICARDO Pompeu
USP, ICMC, Especialista em Ciência de Dados
ricardo.ferreras@gmail.com

Resumo

Este projeto utiliza ciência de dados para desenvolver um sistema de análise de sentimento, aplicável na avaliação da satisfação dos usuários em diversos setores. O sistema, desenvolvido em *Python* e publicado via *API*[1], utiliza técnicas de *NLP*[2] e aprendizado de máquina para transformar grandes volumes de dados textuais em *insights* acionáveis. Resultados mostram uma acurácia de 67,6%, destacando a eficácia do modelo *Random Forest*[3] utilizado. A análise de sentimento permite uma compreensão mais profunda das emoções dos clientes, identifica problemas específicos e áreas de melhoria, e monitora a evolução da percepção dos clientes ao longo do tempo. Quando combinada com outros indicadores de desempenho, proporciona uma visão holística da experiência do cliente, permitindo decisões mais informadas e estratégicas para melhorar continuamente a satisfação dos usuários.

Palavras-chave: ciência de dados, *nlp*, sentimento, *python*, *random forest*

1 Introdução

A análise de sentimento (AS) tem se tornado uma ferramenta essencial na compreensão da satisfação dos usuários, especialmente no contexto digital atual. Este projeto visa desenvolver um *pipeline*[4] robusto para AS, que será publicado como uma *API*[1] acessível para diversas aplicações. A importância desse processo reside na capacidade de transformar grandes volumes de dados textuais em *insights* acionáveis, permitindo que as empresas compreendam melhor as emoções e opiniões dos seus clientes.

O desenvolvimento do *pipeline*[4] envolve várias etapas críticas, desde a coleta e preparação dos dados até a implementação de modelos de aprendizado de máquina, como o modelo *Random Forest*[3]. A publicação da *API*[1] facilita a integração dessa análise em diferentes plataformas e sistemas, tornando os resultados acessíveis e utilizáveis em tempo real.

A AS oferece vantagens significativas na pesquisa de satisfação dos usuários. Ela permite uma compreensão mais profunda das emoções dos clientes, identifica problemas específicos e áreas de melhoria, e monitora a evolução da percepção dos clientes ao longo do tempo. Quando combinada com outros indicadores de desempenho, a análise de sentimento se torna ainda mais poderosa, proporcionando uma visão holística da experiência do cliente e permitindo uma análise mais completa do produto ou serviço oferecido, as empresas podem tomar decisões mais informadas e estratégicas para melhorar continuamente a experiência do cliente.

2 Revisão da Literatura

A AS, também conhecida como mineração de opiniões, é uma área de pesquisa que visa identificar e extrair informações subjetivas de textos escritos em linguagem natural. Segundo Zhang and Liu [2012], essa técnica é fundamental para monitorar a reputação de produtos e marcas, bem como para entender as opiniões dos consumidores. Gomes et al. [2017] realizaram uma revisão sistemática sobre a evolução técnica da análise de sentimento, destacando os métodos mais utilizados e os contextos de aplicação. Melhado and Rabot [2021] exploraram a análise de sentimento como uma ferramenta não apenas de detecção de afetos, mas também de produção emocional, destacando seu papel na psico política e na sociedade de controle.

3 Metodologia

Para o desenvolvimento e publicação desse projeto, utilizamos uma metodologia que integra diversas ferramentas e bibliotecas de software. Primeiramente, empregamos a linguagem de programação *python* devido à sua versatilidade e ampla adoção na comunidade científica. Para a construção e avaliação de modelos de aprendizado de máquina, utilizamos a biblioteca *scikit-learn*, que oferece uma vasta gama de algoritmos e ferramentas para tarefas como classificação, regressão e *clustering*. Além disso, para o processamento de linguagem natural, utilizamos a biblioteca *nltk*, que fornece recursos robustos para *to-*

[1] Interface de Programação de Aplicações, conjunto de padrões, ferramentas e protocolos que permite a comunicação entre diferentes sistemas e aplicações.

[2] Processamento de Linguem Natural

[3] As árvores de decisão representam uma das formas mais simplificadas de um sistema de suporte à decisão

[4] série de etapas ou processos

[5] refere-se ao processo de conversão de uma sequência de texto em partes menores, conhecidas como *tokens*. Esses *tokens* podem ser tão pequenos quanto caracteres ou tão longos quanto palavras

[6] uma das várias técnicas de normalização de texto que converte dados de texto brutos em um formato legível para tarefas de processamento de linguagem natural

[7] técnica essencial no campo do Processamento de Linguagem Natural. Ela busca simplificar a análise textual ao normalizar palavras para sua forma base

kenização[5], *stemming*[6], *lematização*[7], e análise sintática. Essa combinação de ferramentas permite uma abordagem eficiente e eficaz para a análise de dados textuais e a construção de modelos preditivos. E para facilitar o acesso e a integração com outras plataformas, desenvolveu-se uma *API*[1]. Utilizou-se o *framework*[8] *flask* para criar *endpoints*[9] que permitem a consulta e a atualização dos dados, bem como a obtenção de recomendações em tempo real.

3.1 Coleta de dados

A coleta de dados é o primeiro passo na AS. Para este projeto, foram coletadas mais de 21 mil anotações de textos provenientes de diversas fontes, como fóruns, blogs e avaliações de produtos. As avaliações foram categorizadas com base em um sistema de estrelas, onde textos com 1 a 2 estrelas são considerados negativos, 3 estrelas são neutros, e 4 ou 5 estrelas são positivos. Essa categorização permite uma análise mais precisa dos sentimentos expressos pelos usuários, facilitando a identificação de padrões e tendências nas opiniões dos clientes.

3.2 Preparação

A preparação dos dados envolve desde a limpeza até a *tokenização*[5] dos textos. Isso inclui a remoção de *stop words*[10], a correção de erros ortográficos, a normalização de termos e por fim a *tokenização*[5]. Essas etapas são cruciais para garantir que os dados estejam prontos para serem analisados pelos modelos de aprendizado de máquina.

1. Limpeza de dados: Etapa para remover dados irrelevantes ou inconsistentes.
2. Normalização dos textos: Converte todo o texto para um formato padrão, como transformar todas as letras em minúsculas.
3. Remoção de *stop words*[10]: Elimina palavras comuns que não agregam significado, como "e", "de", "a".
4. Correção de erros ortográficos: Corrigir erros de digitação ou ortografia.
5. *Tokenização*[5]: Dividi, separa o texto em unidades menores, como palavras ou frases.

3.3 Modelo e Métricas

Para este projeto, foi desenvolvido um pipeline[4] utilizando o algoritmo *Random Forest*[3]. Este modelo foi escolhido devido à sua robustez e capacidade de lidar com grandes volumes de dados e variáveis. Após o treinamento do modelo com as anotações de textos categorizadas, foi gerada uma matriz de confusão[11] para avaliar a performance do modelo. A matriz de confusão[11] demonstrou a distribuição das previsões corretas e incorretas, permitindo uma análise detalhada da precisão do modelo. A acurácia obtida foi de 67,6%, indicando que o modelo conseguiu classificar corretamente a maioria dos sentimentos expressos nos textos.

A matriz de confusão[11] gerada foi a seguinte:

		Classificação Predita		
		negativo	neutro	positivo
Classificação Real	negativo	1.844 (25,87%)	389 (5,46%)	129 (1,81%)
	neutro	655 (9,19%)	1.317 (18,48%)	447 (6,27%)
	positivo	239 (3,35%)	448 (6,29%)	1.660 (23,29%)

Tabela 2: Matriz de confusão

Além da acurácia, ver tabela 3, outras métricas foram utilizadas para avaliar a performance do modelo em um contexto de classificação multiclasse, incluindo precisão, *recall* e *F1-score*. Essas métricas ajudam a determinar a eficácia do modelo em classificar corretamente os sentimentos expressos nos textos, proporcionando uma visão mais completa da performance do modelo.

- Acurácia (*Accuracy*): Mede a proporção de previsões corretas em relação ao total de previsões. É

[8] estrutura de suporte que serve como base para o desenvolvimento de software

[9] ponto de acesso específico onde as solicitações de *API*[1] são recebidas e processadas

[10] também conhecidas como palavras funcionais ou palavras vazias. São termos de reduzida contribuição semântica ou nocional, que servem prioritariamente para estabelecer relações entre outros vocábulos.

[11] Uma matriz de confusão (ou matriz de erro), tabela 1, é um método de visualização para resultados de algoritmo classificador. Ela contém quatro elementos principais:

- Verdadeiro Positivo (VP): O modelo previu corretamente a classe positiva.
- Falso Positivo (FP): O modelo previu a classe positiva incorretamente (falso alarme).
- Verdadeiro Negativo (VN): O modelo previu corretamente a classe negativa.
- Falso Negativo (FN): O modelo previu a classe negativa incorretamente (erro de omissão).

		Previsão		
		Positivo	Negativo	Total
Teste	Verdadeiro	a	b	$a+b$
	Falso	c	d	$c+d$
	Total	$a+c$	$b+d$	N

Tabela 1: Matriz de confusão

uma boa métrica quando as classes estão balanceadas.

$$\text{Acurácia} = \frac{\text{Previsões corretas}}{\text{Total de previsões}}$$

- Precisão (*Precision*): Mede a proporção de verdadeiros positivos em relação ao total de positivos previstos. Indica a exatidão das previsões positivas.

$$\text{Precisão} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos positivos}}$$

- *Recall* (Sensibilidade): Mede a proporção de verdadeiros positivos em relação ao total de positivos reais. Indica a capacidade do modelo de encontrar todos os positivos reais.

$$\textit{Recall} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos negativos}}$$

- *F1-score*: É a média harmônica entre precisão e *recall*. É útil quando se deseja um equilíbrio entre precisão e *recall*.

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}}$$

Métrica	Valor
Acurácia	0.676
Precisão	0.611
Recall	0.544
F1 Score	0.576

Tabela 3: Métricas de Desempenho

3.4 Exemplo de Uso

GET **/api/sentimento/**
chamada principal da API[1]

Body application/json

```
{
    "texto": "Hoje o dia esta
        excelente para viajar."
}
```

Response application/json

200 ok

```
{
    "Resultado": "positivo",
    "Resultado_Score": {
        "negativo": 0.0452552073,
        "neutro": 0.0855765362,
        "positivo": 0.869168256
    }
}
```

4 Resultados

A análise de sentimento oferece várias vantagens além da pesquisa de satisfação dos usuários tradicional. Primeiramente, permite uma compreensão mais profunda das emoções dos clientes, indo além das métricas tradicionais. Também, facilita a identificação de problemas específicos e áreas de melhoria, permitindo uma resposta mais rápida e eficaz por parte das empresas. Por fim, a análise de sentimento pode ser utilizada para monitorar a evolução da percepção dos clientes ao longo do tempo, ajudando as organizações a adaptar suas estratégias de acordo com as mudanças nas expectativas dos usuários.

5 Conclusão

A análise de sentimento é uma ferramenta poderosa que complementa a análise de satisfação dos usuários, proporcionando *insights* valiosos que podem guiar as decisões estratégicas da empresa. Com a evolução contínua das técnicas e modelos de análise, espera-se que essa área de pesquisa continue a crescer e a oferecer novas oportunidades para a melhoria da experiência do cliente.

Referências

F. P Gomes, E. M. Silva, I Teixeira, and P. F. Brito. Análise de sentimentos: Uma revisão sistemática. *Commun. ACM*, pages 24–32, 2017. ISSN 2447-0767. URL `https://ulbra-to.br/encoinfo/edicoes/2017/artigos/analise-de-sentimentos-uma-revisao-sistematica/`.

Felipe Melhado and Jean-Martin Rabot. Análise de sentimentos: da psicométrica à psicopolítica. *Comunicação e Sociedade*, pages 101–118, 2021. ISSN 1645-2089. doi: 10.17231/comsoc.39(2021).2797. URL `https://hdl.handle.net/1822/73511`.

flask. Flask (version 1.1.2) [computer software]. `https://flask.palletsprojects.com`. Accessed: 2024-12-30.

nltk. Natural language processing with python. `https://www.nltk.org/`. Accessed: 2024-12-30.

python. Python language reference, version 3.8. `https://www.python.org`. Accessed: 2024-12-30.

scikit-learn. Scikit-learn: Machine learning in python, version 1.1.0. `https://scikit-learn.org`. Accessed: 2024-12-30.

Lei Zhang and B. Liu. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*, 2012. doi: 10.1007/978-3-031-02145-9. URL https://api.semanticscholar.org/CorpusID:38022159.