PROJETO EM CIÊNCIA DE DADOS – RELATÓRIO FINAL

(Anexar este arquivo à entrega final do projeto)

SEMESTRE	2023/1
PROJETO	Monitor de tendências da moda
COMPONENTES DO GRUPO	Gustavo da Nóbrega Silva Lorenzo Corrêa Lazzarotto Thiago de Almeida Macedo Vitor Fuentes Ferreira Pires

Breve descrição do problema

Sendo o que segue, a descrição básica do projeto: "O monitor de tendências de moda incremental dará indicativos do que o mercado *fashion* trará nas próximas temporadas, viabilizando *insights* de negócio e auxiliando na tomada de decisão de criação de estilo da empresa. Visa ser uma ferramenta de predição com escopo expansível. Ademais, o conjunto de dados captado auxiliará na modelagem de outros problemas internos, sendo um ponto inicial comum para outras propostas." O grupo entende o projeto como um processo de entrada de uma sequência, no qual ele indicará características que estão sendo tendências no mercado fashion, assim, deve-se caracterizar a popularidade de diferentes estilos e features para alcançar o objetivo proposto, auxiliando em futuras etapas dentro da empresa.

Breve descrição da solução proposta

Como solução para o problema mencionado, o grupo propõem:

Criar uma forma de possibilitar que designers de moda possam acompanhar quais estilos de roupas e acessórios estão surgindo nas redes sociais e como esses estilos são recebidos pelo público. O projeto é composto por uma pipeline que pega dados relacionados a postagens no instagram, incluindo a foto, o número de likes e uma amostra dos comentários. Com esses dados, as roupas são segmentadas e agrupadas, utilizando uma combinação de aprendizado profundo e aprendizado não-supervisionado, enquanto os comentários são classificados como positivos ou negativos. No final, uma dashboard é criada para que os grupos, cada um representando um estilo único de moda, possam ser analisados e comparados.

Fases da Metodologia CRISP-DM

Entendimento do negocio:

Tarefas	Conclusão	Observação
Traçar os objetivos	100%	Foi estudado o problema e traçado o objetivo com maior geração de valor, na

		visão do grupo.
Situação do problema	100%	Foi feita uma análise dos recursos iniciais e como utilizá-los/incrementá-los.
Objetivos da mineração de dados	100%	O grupo utilizou um website chamado Apify para adquirir os dados relacionados às postagens.
Planejamento do projeto	100%	Nas primeiras semanas foi traçado um documento com as etapas do projeto

Entendimento dos dados:

Tarefas	Conclusão	Observação
dado inicial	100%	O dado inicial foi fornecido pelo cliente e avaliado pelo grupo, sendo optado por usar outra fonte.
Descrição dos dados	100%	Foi montado um dataset contendo: id do post, número de likes, número de comentários, uma amostra de 20 comentários e a foto do post.
Exploração do dado	100%	O dataset foi explorado e tratado para solucionar o problema
Qualidade	100%	O dataset ficou funcional para o problema, tendo dados tratados e prontos para serem executados na pipeline.

Data preparation

Tarefas	Conclusão	Observação
Seleção dos dados	100%	Dados selecionados pela interface interativa do site Apify.

Limpeza dos dados	100%	Foram removidos posts sem curtidas ou sem comentários.
Construção de novos dados	100%	Foi feita a tokenização dos textos e a substituição dos emojis pela sua descrição. No lado das imagens, elas foram segmentadas para conter somente a peça de roupa e foram reduzidas a um vetor de dimensionalidade 100 por um autoencoder, modelo de aprendizado profundo.
Integrar novos dados	100%	Foram utilizados os dados do site Apify.
Formatação	–	
Descrição do novo dataset	100%	Foram criados dois datasets, um com uma linha para cada comentario tokenizado e outro para manter as informações sobre cada post como curtidas e comentários. As imagens ficaram em uma pasta exclusiva.

Modeling:

Tarefas	Conclusão	Observação
técnicas	100%	Foram usados diversos modelos de aprendizado de máquina: - YOLO + SAM: segmentação das roupas nas imagens - autoencoder convolucional: criação de um espaço latente que representa uma versão de menor dimensão da imagem original - PCA: algoritmo para diminuir ainda mais a

		dimensionalidade do espaço latente. - KMeans: algoritmo não supervisionado utilizado para agrupar imagens que contém roupas semelhantes - Adaboost: algoritmo supervisionado baseado em árvores e utilizado para classificar os comentários como positivos ou negativos.
teste	100%	
construção	100%	Os algoritmos foram criados através de uma combinação de bibliotecas do python como sklearn, keras e ultralytics. Cada modelo tem seu próprio script.
avaliação	100%	O resultado dos modelos ficou satisfatório. A combinação YOLO + SAM obteve bons resultados, cometendo erros em uma pequena quantidade de imagens. O autoencoder ficou bom, as imagens recriadas ficaram muito parecidas com as originais, mesmo reduzindo seu tamanho de 190 mil valores para 16 mil. O PCA ficou bom, mantendo 95% da variabilidade original. O KMeans não funcionou perfeitamente, criando alguns grupos bons, porém outros sem real significado. Por fim, o Adaboost foi o pior modelo do projeto, obtendo cerca de 75% de f1 score. Isso aconteceu

		devido a baixa quantidade de comentários disponíveis.

Avaliação:

Tarefas	Conclusão	Observação
Avaliar os resultados	100%	O dashboard foi criado, possibilitando a avaliação de cada grupo de roupas gerado.
Rever o processo	100%	O processo foi revisto, o código foi otimizado e melhorado.

Indicação de completude

A ideia original do projeto foi completamente implementada. A pipeline funciona sem problemas e tudo é criado como previsto. Todos os modelos são treinados e salvos para uso posterior.

Autocrítica

O grupo começou o trabalho de forma rápida, conseguindo compreender seus objetivos e traçar uma solução adequada. Foi montada uma pipeline teórica e feito um roteiro de implementação. Baseando-se nesse roteiro, o grupo conseguiu cumprir os prazos estipulados, se atrasando um pouco na parte de modelagem, por conta de possuir outras demandas para serem entregues. Foi decidido separar um tempo para estudos, aprendendo sobre modelos de deep learning e implementação de projeto, como versionamento. Além disso, houve um problema com os dados iniciais, causando a necessidade do grupo de rever a pipeline inicial. Em relação ao CRISP-DM, o grupo acredita ter seguido os passos propostos pelo modelo. Como previsto na disciplina, nem todas as etapas do processo tinham de ser implementadas, então foram realizadas somente as necessárias. Avaliando o ritmo do grupo, acreditamos que conseguimos progredir de forma constante, mesmo que com contratempos no meio tempo. A ideia original do projeto foi dada como impossível de ser realizada, até que o website Apify foi descoberto. No fim, o grupo conseguiu criar algo que apresente significado e embora não seja recomendável aplicar os resultados obtidos no contexto do negócio, foi elaborado um "esqueleto" que precisa de apenas de alguns ajustes para ser algo realmente aplicável. No fim, o grupo conseguiu desenvolver um projeto extremamente complexo e satisfatório para nosso nível de conhecimento.

-x-