Este trabalho tem por objetivo desenvolver e avaliar os seus conhecimentos e habilidades em relação a aplicação das etapas do processo de KDT (Knowledge Discovery in Texts) estudadas em aula: seleção, limpeza, transformação, mineração e interpretação de dados. Trata-se, portanto, de um **projeto de mineração de textos**.

Você é um cientista de dados que trabalha no setor de análise de aplicativos da Google Play Store. A empresa solicitou a você que criasse um mecanismo de classificação automática de sentimentos dos usuários acerca dos 10 aplicativos mais populares da loja. Os aplicativos são:

- Shopee
- SHEIN
- TikTok Lite
- Nubank
- Instagram
- Photo & File Detect
- Whatsapp Messenger
- Canva: Desenho Fotos e Vídeos
- CapCut - Editor de Vídeos
- Gov.br

Para tanto, foram coletados 300 comentários de usuários para cada um dos aplicativos supracitados, totalizando 3.000 comentários. Para cada comentário, avaliadores humanos o classificaram em uma polaridade (positivo ou negativo) e em uma emoção de acordo com as emoções básicas de Ekman e Cordaro (2011) que são:

- **Felicidade:** Representa uma experiência gratificante e positiva para o usuário ao usar o aplicativo;
- **Surpresa:** Indica uma reação inesperada e positiva do usuário durante a utilização do aplicativo;
- **Tristeza:** Reflete o sentimento de descontentamento ou a ausência de uma característica desejada no aplicativo, mas mantém uma apreciação geral pelo aplicativo;
- **Neutro:** Caracteriza a neutralidade na avaliação, com a ausência de manifestações emocionais;
- **Medo:** Sinaliza a presença de medo ou insegurança por parte do usuário relacionado ao uso do aplicativo;
- **Nojo:** Descreve uma sensação desagradável causada por uma característica do aplicativo, prejudicando a experiência do usuário;
- **Raiva:** Atribuída quando há um nível notável de agressividade na avaliação do usuário.

A tabela a seguir apresenta a relação dos dois atributos alvo, isto é, a emoção e a polaridade associada. Nela, observa-se que as emoções "felicidade, surpresa e tristeza" estão associadas a uma polaridade positiva. Já as emoções "medo, nojo e raiva" estão ligadas à polaridade negativa.

Polaridade	Positivo			Neutro	Negativo		
Emoção	Felicidade	Surpresa	Tristeza	Neutro	Medo	Nojo	Raiva

Reiterando, você deve desenvolver um mecanismo para classificar (prever) a emoção e a polaridade dos comentários. Você possui um código base como ponto de partida chamado `Base_Trabalho.ipynb` que faz a predição apenas da emoção utilizando o algoritmo NaiveBayes e que faz a estimação da performance desse classificador. Os comentários estão representados utilizando o modelo *Bag of Words*. Ao final, é realizada a classificação de um comentário aleatório, isto é, que não está na base de treinamento, retirado da Play Store em 2024, condizendo com alguma das emoções de Ekman.

O que você deve fazer

- Representar a mesma base de comentários, utilizando o modelo TF-IDF e testar novamente a estratégia de predição usando o algoritmo NaiveBayes do código base. Faça isso novamente utilizando o algoritmo KNN. Ainda, você deve implementar alguma rede neural com algum embedding (Word2Vec, GloVe, Doc2Vec). O que é isso? Pesquise, descreva o que são e escolha o que melhor se adequar aos dados disponíveis e ao algoritmo de rede neural escolhido para PLN. Por fim, use modelo pré-treinados avançados como o BERTimbau e o Llamma. Não use APIs como a do ChatGPT. Dica: pesquise como utilizar esses modelos do Hugging Face.
    - BoW + NaiveBayes (já feito)
    - TF-IDF + NaiveBayes
    - BoW + KNN
    - TF-IDF + KNN
    - Rede neural com embedding
    - BERTimbau
    - LLama
- Analisar os resultados obtidos em termos de performance (precisão, revocação, medida-F e matriz de confusão). Você deve usar o método de amostragem mais adequado de acordo com o problema (holdout, cross-validation ou estratificado). Justifique sua escolha!
- Comparar as estratégias de predição em termos de performance e escolher aquela com maior performance para realizar **a predição de três comentários não existentes no conjunto original de treino**. Avalie qualitativamente se o

comentário que você escolheu condiz com a emoção e a polaridade determinadas pelo algoritmo.

- Relatar suas escolhas no próprio arquivo Jupyter. Você deve justificar cada escolha, portanto, não deixe linhas de código sem explicação. Crie células de texto explicativas além das células de código. Elas servirão para o seu próprio aprendizado.
    - Por exemplo: Suponha que BoW+KNN obteve performance menor quando comparado com TF-IDF+KNN. Pergunta-se: Por quê? Vá atrás, pesquise e relate suas descobertas!

O que deve ser entregue

- Um repositório no GitHub que contemple todos os critérios de avaliação. O repositório deve ser criado e compartilhado em aula com o professor (contribuidor): https://github.com/gmlunardi

Critérios de avaliação

Critério	**Nota**
Qualidade do Pré-processamento dos Dados • Poucos dados? Dica para possível melhoria de performance: data augmentation	2,0
Escolha e Implementação dos Modelos • Explicação das formas de representação (BoW, TF-IDF, embeddings, etc.) • Explicação de como cada algoritmo funciona	3,0
Avaliação de Desempenho dos Modelos • O porquê da escolha do método de amostragem de dados (treino e teste)	2,0
Classificação e avaliação qualitativa de três comentários novos	1,0
Manutenção de Repositório no GitHub • Commits frequentes e relevantes • A manutenção frequente do repositório é requisito para avaliação do trabalho. Um ou poucos commits implicarão na não avaliação do trabalho!	1,0
Interpretação e Discussão dos Resultados • Discussão sobre as vantagens e limitações das abordagens utilizadas • Originalidade e criatividade • Organização do(s) arquivo(s) Jupyter	1,0

Prazo de entrega

01/08/2024, quinta-feira, até às 15:30.
Trabalhos entregues fora do prazo terão sua nota máxima reduzida em 50%.

Referências

EKMAN, P.; CORDARO, D. **What is meant by calling emotions basic. Emotion review**, Sage Publications Sage UK: London, England, v. 3, n. 4, p. 364–370, 2011.